Michele Costa, Luca De Angelis

MODEL SELECTION IN HIDDEN MARKOV MODELS: A SIMULATION STUDY

Quaderni di Dipartimento

Serie Ricerche 2010, n. 7
ISSN 1973-9346

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Scienze Statistiche "Paolo Fortunati"

Model selection in hidden Markov models: a simulation study

Michele Costa, Luca De Angelis

Abstract A review of model selection procedures in hidden Markov models reveals contrasting evidence about the reliability and the accuracy of the most commonly used methods. In order to evaluate and compare existing proposals, we develop a Monte Carlo experiment which allows a powerful insight on the behaviour of the most widespread model selection methods. We find that the number of observations, the conditional state-dependent probabilities, and the latent transition matrix are the main factors influencing information criteria and likelihood ratio test results. We also find evidence that, for shorter univariate time series, AIC strongly outperforms BIC.

Key words: Model selection procedure, Hidden Markov model, Monte Carlo experiment, information criteria, likelihood ratio test.

1 Introduction

Model selection procedures are a challenging topic in statistical literature and represent an essential step in hidden Markov model estimation. In the framework of hidden Markov models (HMM), model selection plays a prominent role since it corresponds to the choice of the number of latent states, denoted as m, of the unobserved Markov chain underlying the observed data. The number of states should be chosen in order to enable the model to account for the dynamic pattern and the covariance structure of the observed time series. In particular, when using HMM for exploratory purposes, the choice of m is crucial, since in many empirical developments of HMM no clue about its value is available.

Michele Costa,
Dipartimento di Scienze Statistiche, Universita' di Bologna, Italy, e-mail: michele.costa@unibo.it

Luca De Angelis,
Dipartimento di Scienze Statistiche, Universita' di Bologna, Italy, e-mail: l.deangelis@unibo.it

Although model selection has been deeply analyzed for both HMM and other mixture model types, this topic is still an unresolved methodological issue. At the moment, there is not one commonly accepted statistical indicator for deciding the number of latent states of the unobserved Markov chain. Furthermore, the existing studies on this topic provide contrasting findings about the reliability and the accuracy of model selection instruments for HMM.

In the context of finite mixture models, different studies show that Bayesian information criterion (BIC) has a satisfactory behaviour (Fraley and Raftery, 2002; Nylund et al., 2007). However, Fonseca (2008) shows that its performance is not significantly different from the one of other criteria. On the contrary, Akaike information criterion (AIC) is found to have an adequate behaviour for more complex models (Lin and Dayton, 1997) or when the sample size is small (Lukociene and Vermunt, 2010). Furthermore, some authors suggest other criteria for dealing with the model selection issue. For example, Chen and Kalbfleisch (1996) develop a penalized minimum-distance method that provides, under certain conditions, a consistent estimate of the number of components. Moreover, Dias (2006) indicates the AIC3 as the best information criterion for defining the number of classes in latent class analysis. Finally, consistent Akaike information criterion (CAIC) has been proved to have a similar performance with respect to BIC (Lin and Dayton, 1997; Lukociene and Vermunt, 2010).

In HMM framework, the literature dedicated to the issue of model selection is also quite extensive. The consistent identification of the number of latent states, i.e. the order of the unobserved Markov chain, is a fundamental prerequisite for model parameter estimation (Cappé et al., 2005). However, the statement by MacDonald and Zucchini (1997) that 'in the case of HMM, the problem of model selection (and in particular the choice of the number of states in the Markov chain component model) has yet to be satisfactorily solved' is still valid.

Among the works related to this topic, Kerébin (2000) and Csiszár and Shields (2000) demonstrate that BIC is a consistent estimator of the HMM order. MacKay (2002) extends to HMM the approach of Chen and Kalbfleisch (1996) and shows that, in many cases, the penalized minimum-distance method provides a consistent estimate of the number of hidden states in a stationary HMM. Gassiat and Boucheron (2003) prove almost sure consistency of the penalized maximum likelihood estimator with penalties increasing as a power of the order. Moreover, Celeux and Durand (2008) show that cross-validated likelihood criteria can be easily applied when data consists of several independent sequences but become impractical when dealing with univariate time series.

The major aim of this paper is the analysis of model selection procedures in HMM. In order to evaluate and compare existing proposals, we provide a Monte Carlo study which allows a powerful insight on the behavior of the most widespread information criteria.

Information criteria are developed on a two terms structure. The first term, which can be interpreted as a goodness of fit measure, is based on the likelihood function, which is increasing by m, since adding more latent states always improves the fit of the model. The second term is a penalty which has to be traded off against the quadratic increase in the number p of parameters that have to be estimated. The penalty is usually specified as a function of p only or as a function of both p and the number of observations T.

The order of a HMM is usually chosen considering one (or more) information criterion: in the following we refer to the Bayesian information criterion (BIC), the Akaike information criterion (AIC) and four variants of the latter, the AIC3, the consistent AIC (CAIC), the corrected AIC (AICc) and the AICu. Furthermore, we also evaluate the Hannan-Quinn criterion (HQC) which is usually used as an alternative to AIC and BIC in time series analysis.

Although the asymptotic theory of likelihood ratio tests (LRT) is problematic in both the frameworks of mixture modeling (Titterington et al., 1985) and HMM (Gassiat and Kérebin, 2000), it might be interesting to evaluate their performance according to changes in the HMM specification. The main issue about LRT is due to the fact that the null hypothesis is defined on the boundary of the parameter space and consequently the regularity condition of Cramer on the asymptotic properties of the maximum likelihood estimator is not valid. As consequence, the LRT statistic is not asymptotically distributed as a chi-square. However, likelihood ratio tests are found to be consistent if computed via bootstrap (Nylund et al., 2007) and can be used for dependence structure restrictions of an HMM (Giudici et al., 2000). Furthermore, the approach followed by Gassiat and Boucheron (2003) turns out to be somewhat equivalent to generalized likelihood ratio tests. For these reasons, we propose to investigate the capability of LRT of identifying the true order of an HMM by including also this model selection procedure in our experimental analysis.

There exist many factors which could affect the behavior of different model selection procedures. Obviously, it is very complicated, or even unfeasible, to evaluate in a Monte Carlo experiment all possible factors of perturbation and all different model specifications. Thus, in this paper we focus on univariate HMM with homogenous Markov chain and discrete 'state-dependent' processes, firstly by analyzing the effects related to the number of observations. Furthermore, we also try to evaluate the differences related to the level of uncertainty which characterizes the allocation of the observed data to the latent states. Finally, we take into account the degree of complexity of the latent stochastic process underlying the data by analyzing and comparing more and less complex alternatives for the latent transition matrix.

In Section 2, we introduce hidden Markov models and the different model selection procedures we investigate in our simulation study. In Section 3, we describe the Monte Carlo experiment protocol we used and highlight the design factors we evaluate in our analysis. In Section 4, we summarize the results obtained in the Monte Carlo experiments by comparing the behaviour of different model selection methods and their capability of identifying the true HMM order. Finally, in Section 5 we conclude.

2 Model selection in HMM

Hidden Markov models are used for investigating the dynamic pattern of an observed time series $\{X_t\}_{t \in T}$ by means of one discrete latent process $\{Y_t\}_{t \in T}$ governed by a first-order Markov chain.

HMMs have many areas of applications. For instance, they have been widely used for speech recognition (Levinson et al., 1983), in biostatistics (Leroux and Puterman, 1992), for modeling weather conditions (Hughes and Guttorp, 1994), for analyzing longitudinal data (Maruotti and Rydén, 2009) or in other fields such as finance (Rydén et al., 1998), and marketing (Paas et al., 2007).

A univariate HMM with m latent states and homogenous Markov chain is specified as

$$Pr(X_t = x) = \sum_{i,j=1}^{m} Pr(Y_1 = i) \prod_{t=2}^{T} Pr(Y_t = i | Y_{t-1} = j) \prod_{t=1}^{T} Pr(X_t = x | Y_t = i) = \\ = \sum_{i,j=1}^{m} u_i(1) \gamma_{ij} p_i(x) \quad (1)$$

where $u_i(1)$ denotes the prior initial-state probability, γ_{ij} is the generic element of the $m \times m$ matrix Γ and denotes the latent transition probability of switching from state j at time $t-1$ to state i at time t, and $p_i(x)$ indicates the conditional state-dependent probability.

The order of an HMM is the minimum size of the latent state space Y that can generate the observed series X_t. Thus, we are looking for the smallest integer m such that the distribution $\{X_t\}_{t \in T}$ belongs to M^m with parameter space θ^m (Cappé et al., 2005).

In order to define the value of m, we usually resort to some sort of criterion. A model selection criterion is an estimator of the expected discrepancy between the unknown operating model, f, and the fitted model, $g_{\hat{\theta}}$, where $\hat{\theta}$ is an estimate of the parameter vector θ (Linhart and Zucchini, 1986):

$$E\Delta(f, g_{\hat{\theta}}) = 2E\left[\log f(X) - \log g(X|\hat{\theta})\right]. \quad (2)$$

One of the most widespread and used measure based on maximum likelihood estimation is the Kullback-Leibler discrepancy which focuses on the expected log-likelihood:

$$\Delta_{KL}(f, g_\theta) = -E \log g_\theta(X) = -\int \log g_\theta(X) f(X) dX. \quad (3)$$

A traditional approach to the model selection problem based on maximum likelihood consists of penalizing the fit a model by a measure of its complexity. The penalizing term can depend on the number of parameters of the model and/or the number of observations:

$$\hat{m} = \arg\max_{m} \left[\sup_{g \in M^m} \log g(X|\hat{\theta}) - pen(T,p) \right]. \tag{4}$$

In the case of univariate probability distribution and time series, it is possible to derive an expression for the asymptotic value of the expected discrepancy and also to find an asymptotically unbiased estimator of that value, called asymptotic criterion. Under some regularity conditions (see Linhart and Zucchini, 1996, Appendix A1), the asymptotic criterion is obtained from the empirical discrepancy $\Delta_T(\hat{\theta})$, the sample size T, and a trace term K:

$$C = \hat{E}\Delta(f, g_{\hat{\theta}}) = \Delta_T(\hat{\theta}) + K/T \tag{5}$$

for $T \rightarrow \infty$. The term K can assume a complicated expression. However, in the case of the Kullback-Leibler discrepancy (Equation 3), K reduces to the number of parameter p and the asymptotic (simple) criterion is then

$$C_{KL} = \Delta_T(\hat{\theta}) + p/T. \tag{6}$$

This expression is strictly equivalent to the well-known Akaike information criterion (Akaike, 1974):

$$C_{KL} = AIC/2T \tag{7}$$

where

$$AIC = -2\log L_T(\theta) + 2p \tag{8}$$

and L_T denotes the estimated value of the likelihood function.

Since Akaike's seminal work, many authors have proposed different versions of the AIC. Bozdogan (1994) introduced a more penalized version by replacing the 2 in the penalizing term of Equation 8 with 3:

$$AIC3 = -2\log L_T(\theta) + 3p. \tag{9}$$

Hurvich and Tsai (1989) added to the traditional AIC a further quantity which is function of both the number of parameters p and the sample size T:

$$AICc = AIC + [2p(p+1)]/(T-p-1). \tag{10}$$

McQuarrie et al. (1997) changed the AICc in Equation 10 by summing a further addendum:

$$AICu = AICc + T\log[T/(T-p-1)]. \tag{11}$$

Bozdogan (1987) proposed to refer to the traditional AIC in Equation 8 with a penalizing term, which considers also the sample size T:

$$CAIC = -2\log L_T(\theta) + p(1+\log T). \tag{12}$$

Besides the Akaike information criterion and the family of information criteria which directly derive from the AIC original version, in our analysis we also consider a further criterion, the BIC (Schwarz, 1978), which is frequently used within the HMM framework. BIC is achieved in the Bayesian framework with the following expression:

$$BIC = -2\log L_T(\theta) + p\log T. \tag{13}$$

Finally, we also include the criterion proposed by Hannan and Quinn (1979) which is usually used in time series analysis as an alternative to AIC and BIC:

$$HQC = -2\log L_T(\theta) + 2p\log(\log T). \tag{14}$$

As illustrated in section 1, the traditional model selection approach using likelihood ratio tests is problematic in the context of HMM. However, our aim is to assess the capability of LRT to retrieve the true HMM order and compare its behaviour with respect to the information criteria listed above. Therefore, we evaluate the traditional LRT approach, where we compare the null hyphotesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$ through the following test:

$$LRT = -2\left[\log L_T(\theta_0) - \log L_T(\theta_1)\right]. \tag{15}$$

Likelihood ratio tests and information criteria define the set of methods mainly used for determining the order of an HMM and our aim is to analyze their behaviour in the simulation study presented in the following section.

3 Monte Carlo experiment

We propose to assess the performance of the methods described in section 2 through a Monte Carlo experiment in order to evaluate the effect of different design factors.

Since the Monte Carlo experiment allows the a priori knowledge of the true number of latent states m^*, we are interested in comparing m^* with the number of latent states m suggested by the model selection procedures under investigation.

As the first step of our analysis, we simulate different data sets for one discrete random variable X_t with 3 categories, for $t = 1,...,T$. Each data set is generated by imposing the length of the time series. We propose the values $T = 100$, 500, 1000 and simulate 1000 different data sets for each combination. Furthermore, we also consider $T = 5000$ and 10000 in order to assess the property of consistency of the model selection methods.

Furthermore, we define the probability functions which characterize the HMM. First, we set the number of latent states $m^* = 2$ which denotes the true order of the latent Markov process. Second, we set the conditional probabilities $p_i(x)$, for $i = 1,...,m^*$, which indicate the probability of observing the particular value x of variable X_t, given the membership of the observation to the latent state i. Third, we define the latent transition matrix Γ. We also set the initial state probabilities $u_i(1)$ in order to achieve a stationary HMM (see, Zucchini and MacDonald, 2009)[1].

Once we simulated the different data sets with the features specified above, as the second step of our analysis, we estimate HMMs with different number of latent states $h = 1,...,m,...,C$. Then, we compute the information criteria in Equations 8-14 and test whether the model with h states is non-significantly different from the model with $h+1$ states using the LRT in Equation 15. Hence, for each simulated data set, we compare the values of these model selection methods and evaluate their power of detecting the correct number of states.

Among the many factors which could influence HMM selection procedures, in our analysis we evaluate the performance of the different methods first of all with respect to changes in the number of observations, T, which is crucial for the investigation of the asymptotic properties as well as for the most frequent small sample sizes available in many empirical analyses. Furthermore, we pay a particular attention to the conditional probabilities, $p_i(x)$, which reflect the level of uncertainty characterizing the model classification procedure of the time observations to the latent states. Finally, we consider the latent transition probabilities, γ_{ij}, included in matrix Γ which denote different conditions of persistency in the latent Markov chain.

We are well aware that all simulation results are to be carefully treated and that they do not represent a monolithic one-faced truth. In our framework, for example, even though our simulated models are always identified, we could have more alternative models leading to the same results. However, Monte Carlo experiments can be a precious tool able to enlighten on the behaviour of model selection procedures, which is our purpose.

4 Results

In the following, we summarize the main results of the Monte Carlo experiment which show in detail the behaviour of model selection procedures. We take into consideration three design factors, namely the 'state-dependent' conditional probabilities, the latent transition probabilities, and the number of observations.

4.1 The role of the state-dependent probabilities

In the first analysis, we assess changes in the state-dependent conditional probabilities $p_i(x)$ in Equation 1, by considering an HMM with $m^* = 2$ latent states, prior

[1] Setting the initial state probabilities which allow the HMM to have a stationary distribution is not really necessary in our analysis. Since we consider quite large values of T, the effect of initial probabilities is negligible. This is known as the 'forgetting property' (Cappé et al., 2005).

initial-state probabilities $u_i(1) = (0.75, 0.25)$, and the following latent transition matrix:

$$\Gamma = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}.$$

In particular, we consider four different HMMs characterized by the following conditional probabilities $p_i(x)$, for $i = 1, 2$ latent states and $x = 0, 1, 2$ categories of the random variable X_t:

	i / x	0	1	2
HMM_I:	1	0.9	0.1	0.0
	2	0.0	0.1	0.9
HMM_II:	1	0.8	0.2	0.0
	2	0.0	0.2	0.8
HMM_III:	1	0.7	0.2	0.1
	2	0.1	0.2	0.7
HMM_IV:	1	0.6	0.2	0.2
	2	0.2	0.2	0.6

Table 1 shows the percentage of correct order identification for the information criteria and the likelihood ratio test in the case of different time series length and the four models characterized by the different conditional probabilities $p_i(x)$ described above. For each combination of the four models and $T = 100$, 500, 1000, 5000, we generate 1000 data sets for a total of 16000 simulations. From Table 1, it arises that both the state-dependent probabilities and the number of observations play an important role in the order identification problem. A higher level of uncertainty in the allocation of the observations to the latent states implies a worst performance for the seven information criteria and the likelihood ratio test. When T is low, the performance of the model selection procedures declines very fast. In the case of $T = 100$, the best selection method, the AIC, correctly identifies the order 93.1% and 90.1% of the times, for models HMM_I and HMM_II, respectively. However, the percentage drops to 37.1 for model HMM_III and to only 7.2 if we consider HMM_IV. This sharp decline gradually disappers by increasing the number of observations.

Furthermore, Table 1 shows that the performances of the different criteria are quite heterogenous. When T is low, CAIC and BIC are often unable to detect the true number of states which characterize the latent Markov process. On the contrary, AIC and AICc perform much better, followed by AIC3, AICu, and HQC. On the other hand, when the value of T is large, BIC, CAIC, AIC3, AICu, and HQC perform slightly better than AIC and AICc (see e.g., the case of model HMM_III with $T = 5000$ in Table 1).

Despite the well-known problematic issues about its asymptotic distribution, the LRT statistic shows somewhat satisfactory results which are comparable to the percentages related to AICc.

Table 1 Percentage of correct HMM order identification for different values of T and different conditional probabilities $p_i(x)$

	Model	AIC	AIC3	AICc	AICu	CAIC	BIC	HQC	LRT
$T = 100$	HMM_I	93.1	84.3	91.7	81.1	46.2	62.8	83.8	91.2
	HMM_II	90.1	75.3	87.7	71.0	29.2	46.5	74.5	87.6
	HMM_III	37.1	14.7	30.2	11.8	1.0	2.6	14.2	30.3
	HMM_IV	7.2	1.8	5.6	1.1	0.0	0.0	1.6	5.6
$T = 500$	HMM_I	99.1	100	99.1	100	100	100	100	99.1
	HMM_II	98.7	99.9	99.0	99.9	100	100	100	98.8
	HMM_III	97.4	94.6	97.7	94.3	43.7	58.2	90.5	96.7
	HMM_IV	43.7	19.5	42.5	18.2	0.2	0.5	10.2	36.6
$T = 1000$	HMM_I	99.1	100	99.1	100	100	100	100	97.9
	HMM_II	98.6	100	99.0	100	100	100	100	98.7
	HMM_III	98.1	99.8	98.5	99.8	94.0	97.1	100	98.2
	HMM_IV	74.8	49.9	74.8	48.8	1.1	3.2	29.9	69.9
$T = 5000$	HMM_I	98.1	100	98.2	100	100	100	100	98.2
	HMM_II	97.9	99.8	98.0	99.8	100	100	100	98.2
	HMM_III	98.6	100	98.6	100	100	100	100	98.7
	HMM_IV	98.6	99.9	98.6	99.9	88.4	93.3	99.9	98.8

4.2 The role of the latent transition probabilities

The second design factor we believe it is worth to evaluate consists in changes in the latent transition probabilities γ_{ij} collected in matrix Γ. In particular, we consider five HMMs with $m^* = 2$ latent states and the state-dependent probabilities specified as in previous model HMM_II: $p_1(0) = 0.8$, $p_1(1) = 0.2$, $p_1(2) = 0$, and $p_2(0) = 0$, $p_2(1) = 0.2$, $p_2(2) = 0.8$. The five models are characterized by the following latent transition matrices and the initial-state probabilities required in order that HMMs have stationary distributions:

HMM_A: $u_i(1) = (0.50, 0.50)$ and $\Gamma = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$;

HMM_B: $u_i(1) = (0.75, 0.25)$ and $\Gamma = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$;

HMM_C: $u_i(1) = (0.50, 0.50)$ and $\Gamma = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$;

HMM_D: $u_i(1) = (0.5714, 0.4286)$ and $\Gamma = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$;

HMM_E: $u_i(1) = (0.50, 0.50)$ and $\Gamma = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}$.

For each combination of the five models and T = 100, 500, 1000, we generate 1000 data sets for a total of 15000 simulations.

In Table 2 we compare the behaviour of the information criteria and the LRT statistic for the five models specified above and different number of observations. On the basis of the results reported in Table 2, we can affirm that changes in the latent transition probabilities affect the performance of the model selection procedures under investigation. A transition matrix characterized by lower values of state persistence probabilities γ_{ii}, i.e. the probabilities of remaining in the same state from time $t-1$ to time t, reported on the main diagonal of matrices Γ, complicates the task of model selection procedures. In the case of $T = 100$, the percentage of correct order identification for the best selection method, the AIC, computed for model HMM_A is 99.8% in Table 2. This percentage slightly decreases to 90.1% for HMM_B, declines to 60.5% for HMM_C, and drops to 29.5% and 10.2% for models HMM_D and HMM_E, respectively.

The number of observations plays a fundamental role also in this analysis. Increasing the sample size from 100 to 1000 time observations improves the method's capability of detecting the true HMM order. For example, the percentage of BIC in model HMM_D increases from 0.7% in the case of $T = 100$ to 47.1% when $T = 500$ and to 97.7% for $T = 1000$.

Table 2 shows that the information criteria perform quite differently with respect to alternative structures in the latent transition matrix. When $T = 100$, AIC provides the best results, followed by AICc, AIC3, HQC, and AICu, respectively. Also in this analysis, the performance of BIC and CAIC is unsatisfactory for all models except for HMM_A, i.e. the model characterized by the highest persistence. However, a number of observations equals to 500 and 1000 improves the percentage of correct order identification of these two criteria in all the models we considered. From Table 2, it can be noted that the percentage of BIC and CAIC for models HMM_A, HMM_B, and HMM_C reaches 100% in the case of 1000 observations. However, these two criteria identify the true number of latent states only the 26% and 14% of the times, respectively, for model HMM_E when $T = 1000$, thus showing results which are much worse than the other criteria.

Finally, we stress the fact that performance of LRT is comparable to the one of AICc.

4.3 The role of the number of observations

With the third step of our analysis, we aim to assess the consistency property of both information criteria and likelihood ratio test by estimating an HMM characterized by a moderately complicated structure. The consistent estimation of the HMM order is achieved when the model selection method identifies (almost surely) the true number of latent states, for $T \to \infty$.

We analyze model HMM_III specified above by taking into consideration a different number of observations. In particular, we evaluate the behaviour of model selection procedures when the sample size T is equal to 100, 500, 1000, 5000, and 10000 observations. We generate 1000 data sets for each combination for a total of

Table 2 Percentage of correct HMM order identification for different values of T and different latent transition probabilities γ_{ij}

	Model	AIC	AIC3	AICc	AICu	CAIC	BIC	HQC	LRT
$T = 100$	HMM_A	99.8	99.7	99.8	99.7	96.0	98.2	99.7	99.7
	HMM_B	90.1	75.3	87.7	71.0	29.2	46.5	74.5	87.6
	HMM_C	60.5	28.9	52.7	23.1	1.7	7.0	28.0	52.7
	HMM_D	29.5	9.9	23.7	7.3	0.1	0.7	9.3	23.3
	HMM_E	10.2	0.7	7.0	1.1	0.0	0.0	1.5	6.9
$T = 500$	HMM_A	99.1	100	99.4	100	100	100	100	99.1
	HMM_B	98.7	99.9	99.0	99.9	100	100	100	98.8
	HMM_C	98.9	100	99.1	100	92.3	96.1	100	99.0
	HMM_D	98.1	95.9	98.3	95.4	32.1	47.1	89.4	97.7
	HMM_E	74.7	45.9	74.0	44.5	1.0	3.2	30.2	69.3
$T = 1000$	HMM_A	99.1	100	99.4	100	100	100	100	99.1
	HMM_B	98.6	100	99.0	100	100	100	100	98.7
	HMM_C	97.4	99.9	97.8	99.9	100	100	100	97.4
	HMM_D	98.1	100	98.4	100	92.9	97.7	100	98.2
	HMM_E	96.3	91.0	96.3	90.9	14.1	26.0	79.9	95.1

5000 simulations. We favour model HMM_III as its specification is neither too simple nor too complicated, thus we believe that this model is able to highlight method's consistency.

The results reported in Table 3 show that all the criteria increase their percentage of correct order identification for larger values of T. However, the only criteria which reach and stabilize at 100% are BIC, CAIC, and HQC, thus suggesting consistency. It is worth noting that also AIC3 and AICu reach 100% when the sample size is $T = 5000$, however, for $T = 10000$, their percentages slightly decrease to 99.8%. This could be an evidence of a non-consistent behaviour of both AIC3 and AICu. On the other hand, AIC, AICc, and LRT show an excellent performance for $T = 500$ by exceeding 95% of correct order identification, but, when T increases, they tend to stabilize just below 99%.

Table 3 Percentage of correct HMM order identification for different values of T

	AIC	AIC3	AICc	AICu	CAIC	BIC	HQC	LRT
$T = 100$	37.1	14.1	30.2	11.8	1.0	2.6	14.2	30.3
$T = 500$	97.4	94.6	97.7	94.3	43.7	58.2	90.5	96.7
$T = 1000$	98.1	99.8	98.5	99.8	94.0	97.1	100	98.2
$T = 5000$	98.6	100	98.6	100	100	100	100	98.7
$T = 10000$	98.7	99.8	98.7	99.8	100	100	100	98.9

The consistency of the model selection procedures is further analyzed in Table 4 which reports the mean value of the number of latent states indicated in 1000 simulated data sets for each value of T. Table 4 also allows us to evaluate the tendency of the different methods to under- or over-estimate the true number of latent states, $m^* = 2$. The consistent behaviour of BIC, CAIC and HQC is also confirmed by the results reported in Table 4: these information criteria provide a correct estimation of the HMM order for large sample sizes. Moreover, Table 4 shows that the HMM order estimation using AIC, AICc, and LRT is affected by an error, albeit minimal, of overestimation and that also AIC3 and AICu have a tendency to slightly overestimate the number of latent state when the sample size is 1000 and 10000 observations.

Finally, from Tables 3 and 4, it is possible to observe that the correct estimation of m^* for the Hannan-Quinn criterion is achieved with a smaller sample size than the other criteria: when $T = 1000$, HQC correctly estimates the order of the HMM 100% of the times.

Table 4 Mean value of the number of latent states m for data sets simulated from model HMM_III

	AIC	AIC3	AICc	AICu	CAIC	BIC	HQC	LRT
$T = 100$	1.373	1.149	1.304	1.118	1.010	1.026	1.142	1.305
$T = 500$	2.000	1.946	1.997	1.943	1.437	1.582	1.905	1.993
$T = 1000$	2.019	2.002	2.015	2.002	1.940	1.971	2.000	2.018
$T = 5000$	2.014	2.000	2.014	2.000	2.000	2.000	2.000	2.013
$T = 10000$	2.013	2.002	2.013	2.002	2.000	2.000	2.000	2.011

On the whole, we generated 36000 simulated data sets and analyzed the behaviour of LRT and seven information criteria. By comparing the number of latent states suggested by the different methods with m^* we are able to observe and evaluate the accuracy of the model selection procedures, their robustness with respect to changes in the main design factors and, for increasing values of T, also their consistency.

5 Conclusions

We contribute to the debate on HMM selection procedures by developing a Monte Carlo experiment where the true number of latent states is known a priori. We investigate the reliability and the accuracy of the most widespread model selection methods with respect to some relevant features of HMMs. Among the possible design factors, we focus on the effects of the length of the analyzed time series, the conditional state-dependent probabilities, and the latent transition probabilities.

Our simulation study provides interesting insights about the behaviour of different model selection procedures and their capability of detecting the true order of an hidden Markov model.

First of all, our results show how the time-series length, i.e. the number of observations, the conditional 'state-dependent' probabilities, and the latent transition matrix clearly affect the performance of information criteria and likelihood ratio tests. Our results partially clash with the findings reported in previous studies which show that HMM is considerably more sensitive to changes in the conditional probabilities than in perturbations in the transition probability matrix and the initial distribution of the latent Markov chain (Mitrophanov et al., 2005).

On the whole, AIC seems to outperform the other information criteria, followed by AICc. However, these two criteria are affected by an estimation error, albeit minimal, and tend to overestimate the number of latent states also for large sample sizes.

On the other hand, BIC and CAIC provide the worst performances when the number of observations is limited and the HMM is characterized by a higher level of uncertainty or a lower persistence, strongly underestimating the number of latent states. However, if we increase the sample sizes, these criteria improve their performance and, for large sample sizes, they always identify the order of the HMM.

Furthermore, we also evaluate the criteria behaviour for very large values of T. From our results, it emerges that, besides the BIC, whose consistency is already known in literature, also CAIC and HQC show a consistent behavior. Interestingly, the criterion proposed by Hannan and Quinn needs a more limited number of observations in order to reach and stabilize at 100% of correct order identification and also shows a better overall performance than BIC and CAIC.

Among our results, we find an original contribution to the discussion about selection procedures in HMM. Usually, BIC represents one of the preferred methods but, on the basis of our study, we suggest some caution in the use of this criterion when the number of observations in low.

Moreover, despite its well-known asymptotic theoretical issues, the traditional likelihood ratio test shows a behaviour similar to AICc, thus providing a somewhat surprising satisfactory performance in detecting the true order of the HMM.

The simulation study we propose in this paper provides a simple but rigorous framework able to cover all the possible interesting features of HMMs and to evaluate the performance of the most widespread and used model selection methods. Further studies should consider other design factors and other possible model specifications in order to assess whether our results can be extended to different situations. However, our feeling is that together time series length, latent transition probabilities, and uncertainty level represented by conditional probabilities span a wide set of alternatives in HMMs.

References

Akaike H. (1974). A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, 19(6), 716-723.

Bozdogan H. (1987). Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions. *Psychometrika*, 52(3), 345-370.

Bozdogan H. (1994). Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Informational Measure of Complexity. In: Proceedings of the First US/Japan Conference on the Frontiers of statistical modeling: An informational approach vol. 2, Kluwer Academic Publishing, Boston, 69-113.

Cappé O., Moulines E., Rydén T. (2005). *Inference in Hidden Markov Models*, New York: Springer.

Celeux G., Durand J.B. (2008). Selecting Hidden Markov Model State Number with Cross-Validated Likelihood. *Computational Statistics*, 23, 541-564.

Chen J., Kalbfleisch J.D. (1996). Penalized Minimum-Distance Estimates in Finite Mixture Models. *The Canadian Journal of Statistics*, 24, 167-175.

Csiszár I., Shields P. (2000). The Consistency of the BIC Markov Order Estimator. *Annals of Statistics*, 28, 1601-1619.

Dias J.G. (2006). Latent Class Analysis and Model Selection. In: M. Spiliopoulou et al. (Eds.), From Data and Information Analysis to Knowledge Engineering Proceedings of the 29th Annual Conference of the Gesellschaft fr Klassifikation e.V. University of Magdeburg, March 911, 2005, 95-102.

Fonseca J.R.S. (2008). The Application of Mixture Modeling and Information Criteria for Discovering Patterns of Coronary Heart Disease. *Journal of Applied Quantitative Methods*, 3(4), 292-303.

Fraley C., Raftery A.E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of American Statistical Association*, 97, 611-631.

Gassiat E., Boucheron S. (2003). Optimal Error Exponents in Hidden Markov Models Order Estimation. *IEEE Transactions on Information Theory*, 49(4), 964-980.

Gassiat E., Kérebin C. (2000). The Likelihood Ratio Test for the Number of Components in a Mixture with Markov Regime. *ESAIM Probability and Statistics*, 4, 25-52.

Giudici P., Rydén T., Vandekerkhove P. (2000). Likelihood-Ratio Tests for Hidden Markov Models. *Biometrics*, 56, 742-747.

Hannan E.J., Quinn B.G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, B*, 41, 190-195.

Hughes J.P., Guttorp P. (1994). A Class of Stochastic Models for Relating Synoptic Atmospheric Patterns to Regional Hydrologic Phenomena. *Water Resources Research*, 30, 1535-1546.

Hurvich C.M., Tsai C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76(2), 297-307.

Kerébin C. (2000). Consistent estimation of the order of mixture models. *Sankhya*, 62, 4966.

Leroux B.G., Puterman M.L. (1992). Maximum-Penalized Likelihood Estimation for Independent and Markov-Dependent Mixture Models. *Biometrics*, 48, 545-558.

Levinson S.E., Rabiner L.R., Sondhi M.M. (1983). An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *The Bell System Technical Journal*, 62, 1035-1074.

Lin T.H., Dayton C.M. (1997). Model Selection Information Criteria for Non-Nested Latent Class Models. *Journal of Educational and Behavioral Statistics*, 22(3), 249-264.

Linhart H., Zucchini W. (1986). *Model Selection*, New York: Wiley.

Lukociene O., Vermunt J.K. (2010). Determining the Number of Components in Mixture Models for Hierarchical Data. In: A. Fink et al. (Eds.), Advances in Data Analysis, Data Handling and Business Intelligence Proceedings of the 32nd Annual Conference of the Gesellschaft fr Klassifikation e.V., Joint Conference with the British Classification Society (BCS) and the Dutch/Flemish Classification Society (VOC), Helmut-Schmidt-University, Hamburg, July 16-18, 2008, 241-249.

MacDonald I.L., Zucchini W. (1997). Hidden Markov and Other Models for Discrete-Valued Time Series. *Monographs on Statistics and Applied Probability*, 70, London: Chapman & Hall.

MacKay R.J. (2002). Estimating the Order of a Hidden Markov Model. *The Canadian Journal of Statistics*, 30(4), 573589.

Maruotti A., Rydén T. (2009). A Semiparametric Approach to Hidden Markov Models Under Longitudinal Observations. *Statistics and Computing*, 19, 381-393.

McQuarrie A., Shumway R., Tsai C.-L. (1997). The Model Selection Criterion AICu. *Statistics & Probability Letters*, 34, 285-292.

Mitrophanov A.Y., Lomsadze A., Borodovsky M. (2005). Sensitivity of Hidden Markov Models. *Journal of Applied Probability*, 42, 632-642.

Nylund K.L., Asparouhov T., Muthén B.O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling*, 14 (4), 535-569.

Paas L.J., Vermunt J.K., Bijmolt T.H.A. (2007). Discrete Time, Discrete State Latent Markov Modelling for Assessing and Predicting Household Acquisitions of Financial Products. *Journal of the Royal Statistical Society A*, 170, 955-974.

Rydén T., Teräsvirta T., Åsbrink S. (1998). Stylized Facts of Daily Return Series and the Hidden Markov Model. *Journal of Applied Econometrics*, 13, 217-244.

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

Titterington D.M., Smith A.F.M., Makov U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester, UK: Wiley.

Zucchini W., MacDonald I.L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.