Selecting Hidden Markov Model State Number with Cross-Validated Likelihood

Gilles Celeux * Jean-Baptiste Durand †

April 27, 2007

Abstract: The problem of estimating the number of hidden states in a hidden Markov model is considered. Emphasis is placed on cross-validated likelihood criteria. Using cross-validation to assess the number of hidden states allows to circumvent the well documented technical difficulties of the order identification problem in mixture models. Moreover, in a predictive perspective, it does not require that the sampling distribution belongs to one of the models in competition. However, computing cross-validated likelihood for hidden Markov models for which only one training sample is available, involves difficulties since the data are not independent. Two approaches are proposed to compute cross-validated likelihood for a hidden Markov model. The first one consists of using a deterministic half-sampling procedure, and the second one consists of an adaptation of the EM algorithm for hidden Markov models, to take into account randomly missing values induced by cross-validation. Numerical experiments on both simulated and real data sets compare different versions of cross-validated likelihood criterion and penalised likelihood criteria, including BIC and a penalised marginal likelihood criterion. Those numerical experiments highlight a promising behaviour of the deterministic half-sampling criterion.

Keywords: *Hidden Markov Models, Model Selection, Cross-Validation, Missing Values at Random, EM Algorithm*

*Inria Futurs, Orsay, Dept. de mathématiques, Bâtiment 425, Université Paris-Sud, 91405 Orsay Cedex, France

†Laboratoire Jean Kuntzmann – Inria – Grenoble Universités, 51 rue des Mathématiques, B.P.53, 38 041 Grenoble cedex 9, France, Tel. +33 4 76 63 57 09 / fax +33 4 76 63 12 63, email : Jean-Baptiste.Durand@imag.fr

1 Introduction

Finite mixture distributions provide powerful models for statistical learning (see McLachlan and Peel, 2000). In the context of Signal Processing, hidden Markov models (HMM), which are finite mixture models with Markov regime, have been widely used for modelling time series with homogeneous zones, where the observed data have a similar distribution, associated to the states of an unobserved finite state Markov process. For instance, HMM models have been used successfully in various applications such as Speech Recognition (see for instance Rabiner, 1989) or DNA sequence analysis (see for instance Churchill, 1989). For a review on hidden Markov models and their applications, see for example Ephraïm and Merhav, 2002). An important but difficult question to be solved when dealing with a finite mixture model is to choose a relevant number of components or a number of hidden states in the HMM context. Many criteria or procedures have been proposed to answer this open question (see McLachlan and Peel 2000, chapter 6 for a recent state of the art). The main problem occurring with mixture order identification is that the validity conditions for the likelihood ratio tests and related penalised likelihood criteria do not hold. In particular, the vector parameter λ in $\boldsymbol{R}^r$ for the true number of components is not identifiable in a space of higher dimension $r' > r$. For instance, it has been proved that, for general HMMs, the likelihood ratio statistic is stochastically unbounded even for bounded parameters (Gassiat and Kéribin, 2000). However, for an independent mixture model, for which the observed data are independent, it has been shown that on a practical ground the BIC criterion of Schwarz (1978) has often a satisfactory behaviour (see Roeder and Wasserman, 1997 and Fraley and Raftery, 2002). Moreover, this criterion has been proved to lead to a consistent estimation of the number of components of a mixture model in the independent case when the likelihood of the model is bounded (see Kéribin 2000).

BIC, though, is not proved to be consistent for HMMs in the case of general observation distributions. Furthermore, it has been shown that BIC underpenalise the likelihood when determining the number of change-points in change-point processes, a problem closely related to HMMs (Zhang and Siegmund, 2006). Actually such a possible behavior as been noticed for estimating the number of components in a mixture model (see, for instance,

Biernacki, Celeux and Govaert, 2001)

In the independent case, many other criteria have been proposed and experimented in various situations (see Mclachlan and Peel, 2000 chapter 6). For instance the ICL criterion of Biernacki, Celeux and Govaert (2001) has been shown to be relevant to choose a number of mixture components leading to a clustering structure with the greatest evidence (McLachlan and Peel 2000, chapter 6). Another promising criterion is the cross-validated likelihood criterion which has been proposed and experimented by Smyth (2000) in the case of independent mixtures. An advantage of this criterion is that it seems to avoid some of the theoretical difficulties occurring with penalised likelihood criteria and unrealistic assumptions regarding the distribution of the data. There have been a lot of papers from both theoretical and practical viewpoints on choosing the number of components in an independent mixture model. But, as far as we know, there have been a few practical studies on the choice of the number of hidden states for HMMs despite the fact that some penalised likelihood criteria have been analysed in the HMM context (see Boucheron and Gassiat, 2005 for a review). Cross-validation can easily be applied to HMMs when the training data consist of several independent sequences (as in speech recognition for instance). However, applying cross-validation for HMMs estimated on a single sequence is problematic.

The aim of the present article is twofold. Firstly, several ways are proposed to solve the difficulty of implementing the cross-validated likelihood criterion in the HMM context, where the dependencies between the observations have to be taken into account in a proper way. Secondly, numerical experiments on both simulated and real data sets are proposed to compare various criteria to select the number of hidden states in a HMM model. Those criteria are various versions of the cross-validated likelihood criterion, the well known AIC (Akaike, 1973) and BIC criteria, the ICL criterion and the penalised marginal likelihood criterion proposed and studied in Gassiat (2002) for HMMs.

The article is organised as follows. Section 2 is devoted to the presentation and computation of different versions of the cross-validated likelihood criterion. The computation of some of those criteria involves some difficulties when some of the data are removed at random from the learning set of dependent data. Those difficulties are overpassed in the

present paper. Numerical experiments to compare several criteria, including our designed versions of cross-validated likelihood criteria, are presented in Section 3. Subsection 3.2 is devoted to the presentation of Monte Carlo numerical experiments on simulated data. Subsection 3.3 is devoted to the presentation of a real example concerning heart rate variability data for sleeping neonates. The article is concluded with a short discussion section, and the computational details of Section 2 are given in the Appendix.

2 Choosing the number of states of a HMM using cross-validated likelihood

As mentioned in the introduction, assessing the number of components in mixture models encounters theoretical difficulties. A way to bypass those difficulties is to make use of resampling procedures. For instance, McLachlan and Peel (1997) proposed a parameterised bootstrap procedure to the assessment of the P-value of the likelihood ratio test in testing $K = K_0$ components versus $K = K_1$ components. More recently, Smyth (2000) proposed to choose the number of components in a mixture model for independent data by maximising the cross-validated likelihood. The cross-validation approach is natural in model selection because it aims at estimating the predictive performance of a model. For instance, the cross-validated likelihood provides an out-of-sample estimate of the Kullback-Leibler divergence between the actual distribution of the data and a model distribution (Ripley, 1996). An advantage of the cross-validation approach in model selection is that it does not impose the unrealistic assumption that the actual distribution of the data belongs to one of the models in competition as, for instance, AIC and BIC do (see Ripley, 1996 or Spiegelhalter *et al.*, 2000). In the mixture context, an other advantage of cross-validated likelihood is that it circumvents the above-mentioned technical difficulties encountered with penalised likelihood criteria. In this context, using the cross-validated likelihood and especially the half-sampling likelihood has been proved to be a valuable procedure to assess the number of components in finite mixture modelling (Smyth, 2000). In the HMM context, the implementation of cross-validated procedures for determining an appropriate number of hidden states given the data is facing a specific difficulty: removing observations at random from the training data sequence would break the Markovian dependence

between the states of the hidden Markov process. A simple method would consists of partitioning the original sequence into long contiguous blocks and treating them as independent sequences. However, such a strategy is much likely to produce partial sequences where some states of the Markov process are never reached, which would prevent it to find a sensible number of states. Now if, fortunately, a lot of sequences are available, a straightforward cross-validation approach can be implemented. It would consists of using the cross-validation procedure on the sequences rather the observations. This approach has been used in Robertson *et al.* (2004).

However, in this paper, we consider the more general situation where few sequences of data are available. And, in this section, we describe two novel procedures for computing the cross-validated likelihood in the context of hidden Markov processes.

In the sequel, we consider models for sequential data, *i.e.* hidden Markov chains. However, this study can be easily extended to non-sequential processes if the observed data are conditionally independent given the hidden process, for example to hidden Markov trees.A hidden Markov chain consists of an unobserved state process $\boldsymbol{S} = (S_1, \ldots, S_n)$ with finite values $\{1, \ldots, K\}$ and an observed process $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ such that

1. $\boldsymbol{S}$ is a homogeneous Markov chain, which is supposed stationary and ergodic, with stationary state distribution $\boldsymbol{\pi} = (\pi_i)_{1 \leq i \leq K}$ and with transition probability matrix $\boldsymbol{p} = (p_{ij})_{1 \leq i,j \leq K}$;

2. Given a realisation $\boldsymbol{s} = (s_1, \ldots, s_n)$ of $\boldsymbol{S}$, the Y_t are conditionally independent and Y_t follows a distribution with density $f_{\boldsymbol{\theta}_{s_t}}$ belonging to a parametric family $\{f_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$.

The set of parameters of the hidden Markov chain model, denoted $\boldsymbol{\lambda}$ in the following, is usually estimated with the maximum likelihood method by the EM algorithm (Baum *et al.*, 1970).

2.1 Computing cross-validated likelihood criteria for HMMs

When concerned with computing cross-validated maximum likelihood, parameter $\boldsymbol{\lambda}$ has to be estimated from an incomplete observed sequence, since part of the training data set

has been removed. Furthermore the model assessment requires the computation of the likelihood using the remaining data, which also forms an incomplete sequence. To deal with this difficulty, the first approach we propose is a particular half-sampling procedure. It consists of removing from the original training sequence respectively the odd and even indices. The key point is that the resulting processes are still hidden Markov chains, which can be identified through the usual *forward-backward* recursion of Baum *et al.* (1970), *i.e.* the standard implementation of the EM algorithm for HMMs. Since this implementation is subject to underflow when n is moderately large, we use the smoothing algorithm of Devijver (1985), which is numerically stable. This approach provides an easy, rapid and efficient way to implement half-sampling. It preserves the Markovian structure of the model in a simple way, but imposes a particular form of half-sampling. The second approach we consider deals with the more general situation of v-fold cross-validation, where $d = n/v$ among n data point are removed at random at each of the v steps of the cross-validation procedure. This approach has been proposed by Zhang (1993) for the selection of linear models. It can be dealt with the EM algorithm of Dempster *et al.* (1977) and it leads to derive a *forward-backward* recursion for HMMs that is dedicated to data removed at random.

2.1.1 Half-sampling from the odd and even subchains

Half-sampling is a two-fold cross-validation procedure for which the learning and test data sets, of equal size $n/2$, are usually chosen *at random* in the actual data set. However, the indices of those two sub-samples can be chosen in a deterministic way. If the two sub-samples are designed by considering respectively the odd and even indices of the original data set, supposed to follow a HMM distribution, the resulting processes are still HMMs. More precisely, the theorem below holds.

Theorem 1 Distribution of the odd (resp. even) subchain.
Let $(Y_1, \ldots, Y_{2n})$ *be a hidden stationary and ergodic Markov chain with* K *hidden states and with parameter* $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{p}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. *Then the odd (resp. even) subchain* $(Y_1, Y_3, \ldots, Y_{2n-1}) = (Y_1^{'}, \ldots, Y_n^{'})$ *(resp.* $(Y_2, Y_4, \ldots, Y_{2n}) = (Y_1^{''}, \ldots, Y_n^{''})$*) is a hidden Markov chain with* K *hidden states and with parameter* $(\boldsymbol{\pi}, \boldsymbol{p}^2, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$.

Remark: Note that this result would not hold if the Markov chain was periodic. However, since the Markov chain is supposed to be stationary and ergodic, it cannot be periodic.

Proof: For the odd sequence, the result follows from the fact that the hidden process $(S_1, S_3, \ldots, S_{2n-1})$ is straightforwardly a Markov chain with K states, with initial distribution $\boldsymbol{\pi}$ and with transition probability matrix $\boldsymbol{p}^2$. Using the conditional independence of $(Y_1, Y_3, \ldots, Y_{2n-1})$ given $(S_1, S_3, \ldots, S_{2n-1})$ leads to the expected result. The argument is quite analog for the even sequence. The only difference is that the initial distribution of the even chain is $\boldsymbol{\pi p}$. In the case where the original hidden chain is assumed stationary, both hidden subchains are Markov chains with the same distribution, since in this case $\boldsymbol{\pi p}^2 = \boldsymbol{\pi p} = \boldsymbol{\pi}$.$\square$

Consequently, the parameters of the odd and even HMM are the same. Their parameter can be estimated straightforwardly from the EM algorithm and the loglikelihood can be computed as a by-product of the EM algorithm. In the half-sampling procedure the role of the odd and even subsequences are permuted. When the vector parameter $\boldsymbol{\lambda}$ is estimated using the odd (resp. even) subsequence, its loglikelihood is calculated on the even (resp. odd) subsequence using the *forward* recursion of Baum *et al.* (1970). Thus, the resulting OEHS (for Odd-Even Half-Sampling) criterion to assess the number of hidden states of the HMM model is the sum of those loglikelihoods. Since the dominant term of the complexity for each EM iteration is $2nK^2$, this is also the dominant term of the complexity for this half-sampling procedure.

2.1.2 Multifold cross-validation procedure

The implementation of the multifold cross-validation in the general case amounts to delete at random part of the observations and to estimate the vector parameter $\boldsymbol{\lambda}$ and to compute the likelihood from two different incomplete sequences. The whole training sequence $\boldsymbol{y}$ of length n is decomposed at random into an observed subsequence $\boldsymbol{y}_{\boldsymbol{O}bs}$ and a removed subsequence $\boldsymbol{y}_{\boldsymbol{M}is}$ where $\boldsymbol{O}bs$ represents the subset of $\{1, \ldots, n\}$ corresponding to the observed value indices, and where $\boldsymbol{M}is$ is the set $\{1, \ldots, n\} \backslash \boldsymbol{O}bs$ of the removed value indices. To estimate the parameter $\boldsymbol{\lambda}$ in such context, where two types of missing values exist, namely the hidden states and the removed observations, we resort to the EM

algorithm.

The details of the EM algorithm derivation are given in the Appendix. The main results are the following:

- The E step of the EM algorithm amounts to compute the quantities $P(S_t = j, S_{t+1} = k|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})$ and $P(S_t = j|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})$ for each hidden states j and k, at each time t;

- This can be done using the following *forward-backward* recursion. The *forward* recursion is based on the quantities $\tilde{\alpha}_t(j) = P(\{Y_u = y_u\}_{u\in\boldsymbol{O}bs,u\leq t}, S_t = j)$ and the *backward* recursion is based on the quantities $\tilde{\beta}_t(j) = P(\{Y_u = y_u\}_{u\in\boldsymbol{O}bs,t<u\leq n}|S_t = j)$. As shown in the Appendix, those recursions amount to replace $f_{\boldsymbol{\theta}_k}(y_t)$ with the value one when Y_t is removed, in the standard *forward-backward* algorithm of Baum *et al.* (1970). Moreover, our recursion has appealing interpretation: if Y_t is removed and if Y_{t-1} and Y_{t+1} are observed, applying equation (10) followed by equation (9) in the Appendix shows that

  $$\tilde{\alpha}_{t+1}(k) = \sum_j p_{jk} \sum_l p_{lj}\tilde{\alpha}_{t-1}(l) f_{\boldsymbol{\theta}_k}(y_{t+1}) = \sum_j \left[\sum_l p_{jk}p_{lj}\right] \tilde{\alpha}_{t-1}(l) f_{\boldsymbol{\theta}_k}(y_{t+1}).$$

  Thus, it can be noticed that the recursion involves the coefficients of matrix $\boldsymbol{p}^2$. This is due to the transition between S_{t-1} and S_{t+1} being ruled by $\boldsymbol{p}^2$ in the Markov chain $\boldsymbol{S}$. In the same manner, it is easily proved by induction that if n' successive y_t are removed between two observed y_t, the *forward* recursion associated with those two observed y_t involves the transition matrix $\boldsymbol{p}^{n'+1}$. This is illustrated in Figure 1.

- The above implementation of the *forward-backward* recursion is subject to underflow when n is moderately large, as discussed in Devijver (1985). This statement generally applies to HMMs, not just to partially observed HMMs. For this reason, we derive in the Appendix an implementation that is a smoothing algorithm inspired by Devijver's algorithm.

  Both implementations can be used to compute the loglikelihood of a given parameter for a partially observed HMM, using equation (11) of the Appendix for the first implementation, and (18) for the second one.

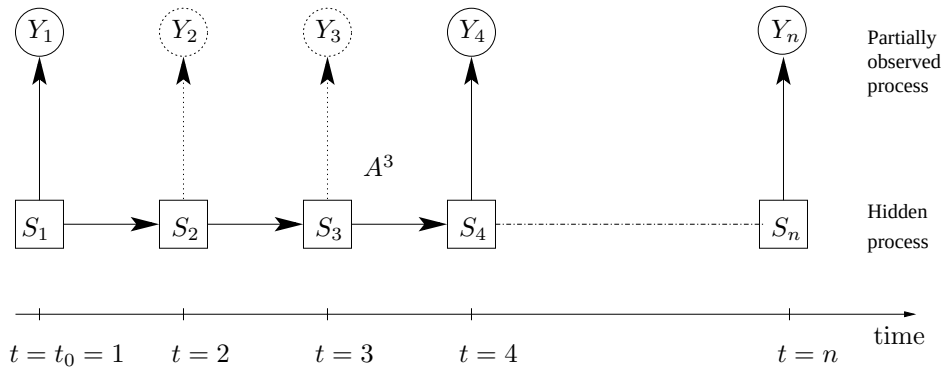


Figure 1: *Deletion of observations with random indices within a hidden Markov chain. If the random variables are deleted from $t = t_0$ to $t = t_0 + n'$, the forward algorithm computes $\tilde{\alpha}_{t+n'+1}$ from $\tilde{\alpha}_t$ by computing $\boldsymbol{p}^{n'+1}$. This is illustrated here for $n' = 2$ and $t_0 = 1$.*

- The M step, given in the Appendix, is slightly different from the standard M step of Baum *et al.* (1970), and has also a direct interpretation in the case of observation distributions in the exponential family.

Remark: As shown in the Appendix, the adaptation of the general multi fold cross-validation scheme to the HMM context appears to be possible without too much difficulty. However, great variability can be expected from such a resampling procedure since useful information to estimate transition probabilities is lost. Moreover, it is quite doubtful that a single run of a multi fold cross-validated procedure would lead to a stable and reliable assessment of the number of hidden states of a HMM. In Durand (2003), numerical experiments highlight the poor behaviour of such single runs of multi fold cross-validation and show that, typically, such procedures have a tendency to overestimate the number of states in a HMM. To reduce this great variability, it is highly recommended to make use of Monte Carlo repetitions of the general multi fold cross-validation scheme. Denoting M the number of Monte Carlo repetitions of the multi fold cross-validation procedure, this leads to compute the cross-validated loglikelihood of a model as the mean of the M cross-validated loglikelihood values derived from the M cross-validation procedures. The EM algorithm detailed in the Appendix can easily be extended to the context of several incomplete sequences. Thus, the cross-validation methodology that consists of deleting observations at random could easily be extended to the case where several sequences are available, though other cross-validation scheme would also be possible.

3 Numerical Experiments

The aim of this section is to give elements on the practical ability of cross-validated likelihood criteria to estimate in a relevant way the number of hidden states in a HMM. In that purpose, the performances of penalised likelihood criteria AIC, BIC, ICL and a Penalised Marginal Likelihood (PML) criterion are compared from numerical experiments with the performances of two cross-validation procedures presented in the previous section to assess the number of hidden states. Before presenting those numerical results, the compared criteria and the experiment conditions are described in the next subsection. In Section 3.2, numerical experiments on simulated data are reported. In Section 3.3, an application of the HMM model to analyse heart rate variability of neonates during sleep is presented.

3.1 Criteria in competition and experiment conditions

The considered cross-validated likelihood criteria are the Odd/Even half-sampling procedure (OEHS) and the Monte Carlo half-sampling procedure (MCHS) described in Section 2.1.2 with half training and half test points.

Penalised likelihood criteria A classical approach to the model assessing problem consists of penalising the fit of a model by a measure of its complexity. A convenient measure of fit is the *deviance* of a model $m \in \mathcal{M}$, which is

$$d(\mathbf{y}) = 2[\ln \mathbf{p}(\mathbf{y}) - \ln \mathbf{p}(\mathbf{y}|m, \hat{\boldsymbol{\lambda}}_m)]$$

where $\mathbf{p}(\mathbf{y})$ denotes the true distribution of the observed data $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{p}(\mathbf{y}|m, \boldsymbol{\lambda}_m)$ denotes the distribution under the model m parameterised with $\boldsymbol{\lambda}_m$, and $\hat{\boldsymbol{\lambda}}_m$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\lambda}_m$. A common way of choosing a penalisation term is to evaluate how large would be the deviance difference on average over learning and test sets of same size. That is, the penalisation would be an estimation of $nD(\boldsymbol{Y}) - E(d(\boldsymbol{Y}))$ where

$$D(\boldsymbol{Y}) = 2E[\ln p(\boldsymbol{Y}) - \ln p(\boldsymbol{Y}|m, \hat{\boldsymbol{\lambda}}_m)]$$

is the expected deviance on a test sample $\boldsymbol{Y}$. Assuming that the data arose from a

distribution belonging to the collection of models in competition and using asymptotic arguments, Akaike (1974) proposed to estimate this difference with $2\nu_m$ where ν_m is the number of free parameters of the model m. This leads to the so-called AIC criterion

$$\mathrm{AIC}(m) = 2 \ln \mathbf{p}(\mathbf{y}|m, \hat{\boldsymbol{\lambda}}_m) - 2\nu_m. \tag{1}$$

An other point of view consists of basing the model selection on the integrated likelihood of the observed data in a Bayesian perspective (see Kass and Raftery, 1995). This integrated likelihood is

$$\mathbf{p}(\mathbf{y}|m) = \int \mathbf{p}(\mathbf{y}|m, \boldsymbol{\lambda}_m)\pi(\boldsymbol{\lambda}_m)d\boldsymbol{\lambda}_m, \tag{2}$$

$\pi(\boldsymbol{\lambda}_m)$ being a prior distribution for parameter $\boldsymbol{\lambda}_m$. A classical asymptotic approximation of the logarithm of the integrated likelihood is the BIC criterion of Schwarz (1978). It is

$$\mathrm{BIC}(m) = \ln \mathbf{p}(\mathbf{y}|m, \hat{\boldsymbol{\lambda}}_m) - \frac{\nu_m}{2} \ln(n). \tag{3}$$

This approximation needs regularity conditions on the likelihoods of the model collection $\mathcal{M}$ and is accurate when the prior distribution $\pi(\boldsymbol{\lambda}_m)$ is centered around the maximum likelihood estimate $\hat{\boldsymbol{\lambda}}_m$ (see Raftery 1995). Notice that it has been argued that this formulation may only be appropriate in circumstances where it was really believed that one and only one of the competing models is in fact true (Bernardo and Smith 1994, chapter 6).

One of the interest of the BIC criterion as compared to the AIC criterion is that under regularity conditions, BIC is expected to be consistent while AIC is not. (For a general theoretical comparison between AIC and BIC, an interesting reference is Yang (2005). The consistency of BIC estimator for HMM number of states is still far from being established (Boucheron and Gassiat, 2005). However, under regularity conditions, Gassiat (2002) has proved the consistency of the following PML criterion to estimate the number of hidden states of a HMM

$$\mathrm{PML}(m) = \sum_{i=1}^{n} \ln p(y_i|m, \hat{\boldsymbol{\lambda}}_m) - \frac{\nu_m}{2} \ln(n), \tag{4}$$

where $p(y_i|m, \hat{\boldsymbol{\lambda}}_m)$ denotes the marginal distribution of y_i for model m. PML as the same form as BIC where the likelihood has been replaced with the pseudo likelihood of $\boldsymbol{\lambda}_m$.

When using a HMM model, it can be advantageous to choose the number of hidden states in order to get the mixture model giving rise to partitioning data with the greatest evidence. With that purpose in mind, Biernacki *et al.* (2001) proposed to approximate the integrated likelihood of the complete data $(\mathbf{y}, \mathbf{s})$ (or integrated completed likelihood), with a BIC-like approximation leading to the criterion

$$\mathrm{ICL}(m) = \ln \mathbf{p}(\mathbf{y}, \hat{\mathbf{s}} \mid m, \hat{\boldsymbol{\lambda}}_m) - \frac{\nu_m}{2} \ln n, \tag{5}$$

where the sequence of missing states has been replaced by its most probable value $\hat{\mathbf{s}}$, for parameter estimate $\hat{\boldsymbol{\lambda}}$, derived through the Viterbi algorithm (see for instance Ephraïm and Mehrav, 2002).

Experiment conditions The simulation and estimation protocols are described hereafter. A procedure now described, inspired from Biernacki *et al.* (2003), is expected to greatly reduce the initial position dependence of the EM algorithm. This procedure makes use of ten different initial parameter values chosen at random. For each initial position, fifty iterations of the stochastic EM *à la* Gibbs algorithm of Robert *et al.* (1993) are first completed. This phase is expected to reduce the dependence of the estimates with respect to the starting position. For each of the ten runs, the parameter value with highest likelihood is used as initial value for the EM algorithm, which is run for fifty iterations. The parameter value with highest likelihood, among the ten different runs, is chosen as initial position for a further long run of the EM algorithm, which is stopped either when 1,000 iterations are completed or when the relative loglikelihood increase is below 10^{-6}. The MCHS criterion is computed using ten Monte Carlo repetitions.

Finally, it is important to note that the cross-validation procedures OEHS and MCHS can be initiated from the MLE derived from the whole observed data sequence. This strategy is expected to reduce the number of iterations required for convergence of the EM algorithm. However, this can lead to suboptimal estimates (compared to random initial values). Moreover, in a cross-validation procedure, the estimates to be assessed are not supposed to be dependent of the samples used to assess its performance. This is why we use the full sequence MLE as an eleventh initial parameter for the algorithm. This mixed strategy empirically has led to the highest likelihood of the final parameter,

compared to those based of either ten random initial values, or only the full sequence MLE. The identification of partially observed HMMs has been performed with new MATLAB routines[1].

3.2 Simulated data

Thirty sequences of length 350 have been simulated, using a two-dimensional stationary Gaussian HMM model with three states, defined by the parameter $\boldsymbol{\lambda}_0$ composed with

$$\boldsymbol{p} = \begin{bmatrix} 0.7 & 0.15 & 0.15 \\ 0.1 & 0.7 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 2.0 \\ 1.9 \end{bmatrix} \quad \boldsymbol{\mu}_3 = \begin{bmatrix} 2.0 \\ -1.9 \end{bmatrix}$$

and the variance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \mathrm{Id}$, where Id refers to the identity matrix. Figure 2 (a) displays the thirty simulated data sets on the same picture. When estimating the Gaussian HMM model, the variance matrices are assumed to be of the form $\gamma\mathrm{Id}$, where the unknown scalar γ is independent from the Markov chain state. In this experiment, the number of hidden states K varies between $K = 1$ and $K = 7$. The frequencies of choosing a K value with each criterion are reported in Table 1.

Criterion	Number of selected hidden states						
	1	2	3	4	5	6	7
MCHS	-	-	**96.67** %	3.33 %	-	-	-
OEHS	-	-	**100.00** %	-	-	-	-
AIC	-	-	**50.00** %	**50.00** %	-	-	-
BIC	-	-	**100.00** %	-	-	-	-
ICL	-	-	**100.00** %	-	-	-	-
PML	-	23.33 %	**76.67** %	-	-	-	-

Table 1: *Frequencies of choosing a number of hidden states for the three-state simulated HMM with the criteria MCHS, OEHS, AIC, BIC, ICL and PML. The frequencies are computed over thirty simulated sequences of length 350.*

It appears that except PML and AIC, all the criteria select the generating model. Criterion PML has a slight tendency to underestimate K while AIC has a marked tendency to select more complex models.

Further experiments, performed using a generating HMM model with five hidden states less separated than those defined by $\boldsymbol{\lambda}_0$, are now presented. Thirty sequences of length

[1]routines available at `http://www-lmc.imag.fr/lmc-sms/Jean-Baptiste.Durand/software.html`

350 have been simulated, using a two-dimensional stationary Gaussian HMM model with five states, defined by the parameter $\boldsymbol{\lambda}_1$ composed with

$$\boldsymbol{p} = \begin{bmatrix} 0.7 & 0.1 & 0.1 & 0.05 & 0.05 \\ 0.05 & 0.8 & 0.05 & 0.07 & 0.03 \\ 0.03 & 0.15 & 0.75 & 0.03 & 0.04 \\ 0.02 & 0.01 & 0.05 & 0.85 & 0.07 \\ 0.04 & 0.02 & 0.02 & 0.02 & 0.9 \end{bmatrix},$$

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 2.7 \\ 2.7 \end{bmatrix} \quad \mu_3 = \begin{bmatrix} 2.7 \\ -2.7 \end{bmatrix} \quad \mu_4 = \begin{bmatrix} 2.0 \\ 1.5 \end{bmatrix} \quad \mu_5 = \begin{bmatrix} 2.0 \\ -1.5 \end{bmatrix},$$

and the variance matrices $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_5 = \mathrm{Id}$. Moreover, for the same model, thirty sequences of length 1,400 have been simulated. Those data sets are displayed in Figure 2 (b) and (c) respectively.

In the estimation procedure, the number of maximal iterations has been increased from 1,000 to 1,500 for the sequences of length 350, and to 2,000 for the sequences of length 1,400. *A posteriori*, this was not necessary since for most sequences, convergence was achieved before 1,000 iterations.

Criterion	Number of selected hidden states						
	1	2	3	4	5	6	7
OEHS	-	-	**50.00** %	26.67 %	16.67 %	6.67 %	-
MCHS	-	-	33.33 %	**40.00** %	23.33 %	3.33 %	-
AIC	-	-	6.67 %	**50.00** %	40.00 %	3.33 %	-
BIC	-	-	**90.00** %	10.00 %	-	-	-
ICL	-	6.67 %	**76.67** %	16.67 %	-	-	-
PML	-	30.00 %	**70.00** %	-	-	-	-

Table 2: *Frequencies in percentage of choosing a number of hidden states for the five-state simulated HMM with the criteria MCHS, OEHS, AIC, BIC, ICL and PML. The frequencies are computed over thirty simulated sequences of length 350.*

From Table 2, it appears that no criterion selects the generating model in the majority of cases, and most criteria favours the three-state solution, which seems reasonable as shown in Figure 2(b). From this set of numerical experiments, only cross-validation criteria and AIC indicate the five hidden state solution as a plausible solution. Moreover, in the present case, AIC seems to have the most satisfactory behaviour, but it is probably favoured here by its tendency to underpenalise the complexity of a model.

Criterion	Number of selected hidden states						
	1	2	3	4	5	6	7
OEHS	-	-	-	10.00 %	**53.33** %	33.33 %	3.33 %
MCHS	-	-	-	-	**83.33** %	13.33 %	3.33 %
AIC	-	-	-	-	**56.67** %	40.00 %	3.33 %
BIC	-	-	-	26.67 %	**73.33** %	-	-
ICL	-	-	10.00 %	**66.67** %	23.33 %	-	-
PML	-	-	**100.00** %	-	-	-	-

Table 3: *Frequencies in percentage of choosing a number of hidden states for the five-state simulated HMM with the criteria MCHS, OEHS, AIC, BIC, ICL and PML. The frequencies are computed over thirty simulated sequences of length 1,400.*

With a larger sequence of length 1400, all the criteria except ICL and PML, select most often the same number of hidden states as the generating model, as it appears in Table 3. The most satisfactory results are obtained with criterion MCHS and to a smaller extend with criterion BIC. Criteria OEHS and AIC have a tendency to overestimate the number of hidden states, contrary to PML, which seems to highly underestimate the number of hidden states, as ICL does to a smaller extend. Actually, as shown in Figure 2(b) and (c), the five mixture components are poorly separated. From this point of view, the behaviour of ICL, which favours no overlapping solutions by its very nature, is sensible. Concerning the behaviour of PML, since the mixture components are overlapping, it seems difficult to diagnosis the presence of five hidden states without taking into account the Markovian dependency between data. However, the PML criterion does not.

To measure the asymptotic ability of the competing criteria to select the right number of hidden states, 30 sequences of length 15,000 have been simulated from the same model.

Criterion	Number of selected hidden states						
	1	2	3	4	5	6	7
OEHS	-	-	-	-	**100.00 %**	-	-
AIC	-	-	-	-	**93.33 %**	6.67 %	-
BIC	-	-	-	-	**100.00 %**	-	-
ICL	-	-	-	-	**100.00 %**	-	-
PML	-	-	-	-	**96.67** %	3.33 %	-

Table 4: *Frequencies in percentage of choosing a number of hidden states for the five-state simulated HMM with the criteria OEHS, AIC, BIC, ICL and PML. The frequencies are computed over 30 simulated sequences of length 15,000.*

The results are represented in Table 4. The procedure MCHS, which is much more CPU time costly, had not been included in those large sample size numerical experiments. It appears from table 4 that OEHS, BIC and ICL selected five-state models in the every case. AIC selected a six-state model for two sequences, and PML selected a six-state model for one sequence (with values of the criteria very close to those of a five-state and of a three-state model) and a five-state model for the other sequences.

Those results tend to highlight a correct behaviour of the three criteria for large samples, although BIC and ICL have a more marked maximum on the number of states of the generating model. It is worth noticing that for relatively small sample sizes, the chosen number of states with most criteria is smaller than the true number of states of the generating model, as it appears from Table 2. This behaviour is no surprising: for small sample sizes, a more parsimonious model can be expected to produce more stable prevision.

3.3 Heart rate variability in sleeping neonates

We consider the problem of heart rate variability analysis in sleeping neonates addressed in Clairambault *et al.* (1992). The signal has been modelled using a Gaussian HMM in Celeux and Clairambault (1992). Their aim was to identify several periods in the neonate sleep, which can be characterised by states of the heart rate variability and can be interpreted as quiet sleep, intermediate sleep and agitated (or so-called active) sleep. The problem of selecting the number of hidden states has not been considered yet, and the authors chose an *a priori* two-state HMM corresponding to calm and active sleep modelling. Though the hypothesis of a Gaussian emission distribution family leads to operational restoration of the periods, this model does not appear to fit the data accurately.

Consequently, we use a transformed signal where the number of heartbeats per second is computed. This transformation induces a loss of information, but leads to an easy segmentation of the signal into biologically interpretable periods. Since in our sample this quantity varies between 1 and 4, we model this discrete signal by a HMM with multinomial emission distributions, and with stationary state distribution. The parameter is estimated using the same procedure as in 3.2, except that three random values are drawn as initial parameters for the EM algorithm. Figure 3 presents the values of each selection criterion

for one sequence of length 5,356.

Both BIC and ICL select a three-state model. The MCHS selects a two-state model. Strictly speaking, the OEHS selects a five-state model, but the value of the criterion for this model is not much different from that for a two-state model, which would actually be selected by a modeller.

The MLE associated with a three-state model is the following:

$$\hat{\boldsymbol{\pi}} = \begin{bmatrix} 0.73 & 0.15 & 0.12 \end{bmatrix}; \hat{\boldsymbol{p}} = \begin{bmatrix} 0.99 & 0.01 & 0.00 \\ 0.07 & 0.27 & 0.66 \\ 0.00 & 0.80 & 0.20 \end{bmatrix}; \begin{array}{l} \hat{\theta}_1 = \\ \hat{\theta}_2 = \\ \hat{\theta}_3 = \end{array} \begin{bmatrix} 0.00 & 0.03 & 0.88 & 0.09 \\ 0.01 & 0.00 & 0.99 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix},$$

which means that $P(Y_t = 1|S_t = 1) = 0.00$, $P(Y_t = 2|S_t = 1) = 0.03$, $P(Y_t = 3|S_t = 1) = 0.88$, $P(Y_t = 4|S_t = 1) = 0.09$, *etc.* Note that the actual frequency of $\{Y_t = 1\}$ is approximately 3.7×10^{-4}.

The MLE associated with a two-state model is:

$$\hat{\boldsymbol{\pi}} = \begin{bmatrix} 0.65 & 0.35 \end{bmatrix}; \hat{\boldsymbol{p}} = \begin{bmatrix} 0.99 & 0.01 \\ 0.02 & 0.98 \end{bmatrix}; \begin{array}{l} \hat{\theta}_1 = \\ \hat{\theta}_2 = \end{array} \begin{bmatrix} 0.00 & 0.03 & 0.89 & 0.08 \\ 0.00 & 0.00 & 0.62 & 0.38 \end{bmatrix},$$

which essentially consists of merging states 2 and 3 of the three-state model. This is also highlighted by the hidden state restoration performed by the Viterbi algorithm providing the globally most likely sequence of hidden states, as shown in Figure 4.

The PML selects a one-state model, and is a decreasing function of the number of hidden states. Such a one-state model is clearly inappropriate to model periods in neonatal sleep.

From a biological viewpoint, two or three stages of sleep are defined in neonates (unlike in adults, who have five stages of sleep): quiet sleep, active sleep, and sometimes intermediate sleep, which is vaguely defined. The HMM model aims at associating the hidden states with stages of sleep. With a three-state model, states 2 and 3 appear to be unstable, as shown by the hidden state restoration on Figure 4 and by the diagonal coefficients of the estimated transition probability matrix. Thus, a sojourn into state 2 or state 3 cannot be interpreted as a well-defined stage of sleep, not even as intermediate sleep. In contrast, the states of the two-state model are stable and are more likely to be interpreted as stages of sleep. Therefore, a two-state model seems to be preferable to a three-state model. In the former model, states 1 and 2 would correspond to quiet sleep

(characterised by slow heart rate) and active sleep (high heart rate), respectively. The three-state model leads to the splitting of state 2 of the two-state model into two new states that are poorly separated. Oscillations in the state process are the consequence of selecting an additional unnecessary state, rather than actual oscillations in the stage of sleep.

4 Discussion

It has been proposed to assess the number of hidden states in a HMM with the cross-validated likelihood. Owing to the dependence of the data sequence occurring with HMM models, the computation of cross-validated likelihood induced some difficulties. Those difficulties have been circumvented using a deterministic half-sampling scheme or solved using a new version of the EM algorithm to estimate the parameter of a HMM with data missing at random. This leads to define two cross-validation criteria of different nature to assess the number of hidden states in a HMM.

Both criteria, the so-called OEHS and MCHS have been compared with penalised likelihood criteria AIC, BIC, PML and ICL from numerical experiments. The conclusions of those numerical experiments are the followings.

- Criteria AIC, BIC and ICL seem to have a behaviour in the HMM context analogous to their behaviour in the independent mixture context (see McLachlan and Peel, 2000, chapter 6). Criterion AIC has a tendency to underpenalise the complexity of a model, ICL favours models that give rise to partitioning the data with the greatest evidence from the hidden states, and BIC seems to perform well if a HMM gives a reasonable representation of the observed process. However, for the heart rate variability data sets, BIC seems to overestimate the complexity of the models, as it could happen for real data sets. It provides a hidden representation of the data in three sets highly less useful than the two-state representation provided by the cross-validated criteria as it appears from Figure 4.

- In the HMM context, only the PML criterion has been proved to be consistent under regularities conditions (Gassiat, 2002). However, it seems that PML converges very

slowly to the optimal solution. Moreover, in practical situations, it seems to have a high tendency to overpenalise the complexity of a HMM model when the sequence length is not very large. Actually, it can be remarked that the maximal pseudo likelihood is always smaller than the maximal likelihood for any HMM. It means that a good fit is less rewarded with the pseudo likelihood than with the likelihood.

- The cross-validation criterion MCHS seems to have a satisfactory behaviour. From our experiments, MCHS may have a slight tendency to select more complex models than BIC does on simulated datasets, but a more parsimonious one on the considered real dataset. Its main disadvantage is that it is much more time consuming than the other criteria. Only a two fold cross-validation criterion has been experimented; however, a Monte Carlo ten (for instance) fold cross-validation criterion would expectedly lead to analogous and more stable performance without producing a deterioration of the CPU time.

- Criterion OEHS can be regarded as a good surrogate to the MCHS criterion. It is highly less time consuming and seems to have analogous performance.

It can be argued that periodic subsampling schemes (as OEHS) will not work when the hidden process is also periodic, which is a very particular case. For such chains, some diagonal coefficients of the transition matrix will be null. Howerer, it is interesting to assess the behaviour of OEHS for diagonal coefficients close to zero. For this reason, further experiments have been performed in Durand (2003), using a three-state model with following transition matrices: matrix $\boldsymbol{p} = \boldsymbol{p}_1$ of Section 3.2, and

$$\boldsymbol{p}_2 = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad \boldsymbol{p}_3 = \begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 0.6 & 0.1 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}.$$

Thirty sequences were simulated. The results are given in Table 5. A less efficient initialisation procedure was used, which explains the differences with Table 1. These results show that the performance of OEHS remains reasonable when the sojourn time into each state is very short.

Transition matrix	$\boldsymbol{p}_1$	$\boldsymbol{p}_2$	$\boldsymbol{p}_3$
Number of selected hidden states	= - / +	= - / +	= - / +
OEHS	**90** 0 / 10	**100** 0 / 0	**100** 0 / 0
MCHS	**73** 0 / 27	**87** 0 / 13	**87** 0 / 13
BIC	**90** 0 / 10	**94** 3 / 3	**97** 0 / 3

Table 5: *Frequencies of choosing a correct (=), too low (-) or too high (+) number of states for three-state simulated HMMs with the criteria OEHS, MCHS and BIC, using three transition matrices and thirty sequences of length 350.*

Acknowledgements

The author are thankful to J. Clairambault and P. Gonçalvès for helpful discussions concerning the application.

Appendix: EM algorithm for HMMs with deleted observations

For any parameter $\boldsymbol{\lambda}$ and any occurrence of the hidden sequence $\boldsymbol{s}$, the completed data loglikelihood of $\boldsymbol{\lambda}$ is defined by

$$\begin{aligned}\ln \mathcal{L}_{\boldsymbol{y}_{\boldsymbol{O}bs},\boldsymbol{y}_{\boldsymbol{M}is},\boldsymbol{s}}(\boldsymbol{\lambda}) &= \sum_{t=1}^{n}\sum_{j=1}^{K} \ln f_{\boldsymbol{\theta}_j}(y_t)\mathbf{I}_{\{s_t=j\}} \\ &+ \sum_{t=1}^{n-1}\sum_{j=1}^{K}\sum_{k=1}^{K} \ln p_{jk}\mathbf{I}_{\{s_t=j,s_{t+1}=k\}} \\ &+ \sum_{j=1}^{K} \ln \pi_j \mathbf{I}_{\{s_t=j\}} \qquad (6)\end{aligned}$$

where $\mathbf{I}_{\{.\}}$ denotes the indicator function. Let $\boldsymbol{\lambda}^{(m-1)}$ be the parameter estimate at iteration $m-1$ of the EM algorithm. The function Q of $\boldsymbol{\lambda}$, to be maximised in the M step of the EM algorithm, is the conditional expectation of the completed data loglikelihood for the current parameter value $\boldsymbol{\lambda}^{(m-1)}$,

$$Q(\boldsymbol{\lambda},\boldsymbol{\lambda}^{(m-1)}) = E_{\boldsymbol{\lambda}^{(m-1)}}[\ln \mathcal{L}_{\boldsymbol{Y}_{\boldsymbol{O}bs},\boldsymbol{Y}_{\boldsymbol{M}is},\boldsymbol{S}}(\boldsymbol{\lambda})|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}].$$

It follows from equation (6) that

$$
\begin{aligned}
Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(m-1)}) &= \sum_t \sum_j E_{\boldsymbol{\lambda}^{(m-1)}}[\ln f_{\boldsymbol{\theta}_j}(Y_t)\mathbf{I}_{\{S_t=j\}}|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}] \\
&+ \sum_t \sum_j \sum_k \ln p_{jk} P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j, S_{t+1} = k|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}) \\
&+ \sum_j \ln \pi_j P_{\boldsymbol{\lambda}^{(m-1)}}(S_1 = j|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}). \qquad (7)
\end{aligned}
$$

This expression can be simplified by noting that if $t \in \boldsymbol{O}bs$, *i.e.* if Y_t is considered as observed,

$$
E_{\boldsymbol{\lambda}^{(m-1)}}[\ln f_{\boldsymbol{\theta}_j}(Y_t)\mathbf{I}_{\{S_t=j\}}|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}] = \ln f_{\boldsymbol{\theta}_j}(y_t) P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}),
$$

and, if $t \in \boldsymbol{M}is$, *i.e.* if Y_t has been removed from the training sample,

$$
\begin{aligned}
&E_{\boldsymbol{\lambda}^{(m-1)}}[\ln f_{\boldsymbol{\theta}_j}(Y_t)\mathbf{I}_{\{S_t=j\}}|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}] \\
&\quad = E_{\boldsymbol{\lambda}^{(m-1)}}[\ln f_{\boldsymbol{\theta}_j}(Y_t)|S_t = j] P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})
\end{aligned}
$$

since, given $S_t = j$, Y_t does not depend on any other random variable, and particularly on the $\boldsymbol{Y}_{\boldsymbol{O}bs}$ process. Thus, equation (7) can be written

$$
\begin{aligned}
Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(m-1)}) &= \sum_{t \in \boldsymbol{O}bs} \sum_j \ln f_{\boldsymbol{\theta}_j}(y_t) P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}) \\
&+ \sum_{t \in \boldsymbol{M}is} \sum_j E_{\boldsymbol{\lambda}^{(m-1)}}[\ln f_{\boldsymbol{\theta}_j}(Y_t)|S_t = j] P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}) \\
&+ \sum_t \sum_j \sum_k \ln p_{jk} P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j, S_{t+1} = k|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}) \\
&+ \sum_j \ln \pi_j P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs}). \qquad (8)
\end{aligned}
$$

E step

The E step of the EM algorithm consists of computing $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(m-1)})$. From (8), this amounts to compute the quantities $P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j, S_{t+1} = k|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})$ and $P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j|\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})$ for each hidden states j and k, at each time t. This can be done using the following *forward-backward* recursion. In the sequel, $P_{\boldsymbol{\lambda}^{(m-1)}}$ is denoted P for the sake of simplicity.

The *forward* recursion is based on the quantities $\tilde{\alpha}_t(j) = P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t}, S_t = j)$. It is initiated at $t = 1$ with

$$
\tilde{\alpha}_1(j) = \begin{cases} P(Y_1 = y_1, S_1 = j) = \pi_j f_{\boldsymbol{\theta}_j}(y_1) & \text{if } 1 \in \boldsymbol{O}bs; \\ P(S_1 = j) = \pi_j & \text{if } 1 \in \boldsymbol{M}is. \end{cases}
$$

Then the $(\tilde{\alpha}_t(j))_j$ are computed inductively for increasing values of t. If $t+1 \in \boldsymbol{O}bs$, $\tilde{\alpha}_{t+1}(k)$ is obtained from the $(\tilde{\alpha}_t(j))_j$ in a standard fashion as in Baum *et al.* (1970), *i.e.*

$$\tilde{\alpha}_{t+1}(k) = \sum_j p_{jk}\tilde{\alpha}_t(j) f_{\boldsymbol{\theta}_k}(y_{t+1}). \tag{9}$$

If $t+1 \in \boldsymbol{M}is$, $\tilde{\alpha}_{t+1}(k)$ is computed by integrating $P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t+1}, S_{t+1} = k)$ over y_{t+1}. From equation (9), this leads to

$$\begin{aligned}\tilde{\alpha}_{t+1}(k) &= \int_{y_{t+1}} \sum_j p_{jk}\tilde{\alpha}_t(j) f_{\boldsymbol{\theta}_k}(y_{t+1})\, dy_{t+1} \\ &= \sum_j p_{jk}\tilde{\alpha}_t(j) \int_{y_{t+1}} f_{\boldsymbol{\theta}_k}(y_{t+1})\, dy_{t+1} = \sum_j p_{jk}\tilde{\alpha}_t(j),\end{aligned} \tag{10}$$

since $\tilde{\alpha}_t(j)$ does not depend on y_{t+1} and since $f_{\boldsymbol{\theta}_k}$ is a probability density function. The likelihood is given by

$$P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs}) = \sum_j \tilde{\alpha}_n(j) \tag{11}$$

The *backward* recursion is based on the quantities $\tilde{\beta}_t(j) = P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, t < u \leq n} | S_t = j)$. It is initiated at $t = n$ with $\tilde{\beta}_n(j) = 1$. Then the $(\tilde{\beta}_t(j))_j$ are computed inductively for decreasing values of t. If $t+1 \in \boldsymbol{O}bs$, $\tilde{\beta}_t(j)$ is obtained from the $(\tilde{\beta}_{t+1}(k))_k$ as in Baum *et al.* (1970), *i.e.*

$$\tilde{\beta}_t(j) = \sum_k p_{jk}\tilde{\beta}_{t+1}(k) f_{\boldsymbol{\theta}_k}(y_{t+1}). \tag{12}$$

If $t+1 \in \boldsymbol{M}is$, $\tilde{\beta}_t(j)$ is computed by integrating $P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, t < u \leq n} | S_t = j)$ over y_{t+1}. From equation (12), this leads to

$$\begin{aligned}\tilde{\beta}_t(j) &= \int_{y_{t+1}} \sum_k p_{jk}\tilde{\beta}_{t+1}(k) f_{\boldsymbol{\theta}_k}(y_{t+1})\, dy_{t+1} \\ &= \sum_k p_{jk}\tilde{\beta}_{t+1}(k) \int_{y_{t+1}} f_{\boldsymbol{\theta}_k}(y_{t+1})\, dy_{t+1} = \sum_k p_{jk}\tilde{\beta}_{t+1}(k),\end{aligned} \tag{13}$$

since $\tilde{\beta}_{t+1}(k)$ does not depend on y_{t+1}.

The above implementation of the *forward-backward* recursion is subject to underflow when n is moderately large, as discussed in Devijver (1985).

For this reason, we recommend the following implementation, which is a smoothing algorithm inspired by Devijver's algorithm. The corresponding recursions are based on

the quantities

$$\alpha_t(j) = P(S_t = j | \{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t})$$

and

$$\beta_t(j) = \frac{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, t < u \leq n} \,|\, S_t = j)}{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, t < u \leq n} \,|\, \{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t})}.$$

Thus

$$\alpha_t(j) = \frac{\tilde{\alpha}_t(j)}{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t})} \tag{14}$$

The *forward-backward* recursion of the smoothing algorithm is the following. For each state k and each time $t > 1$, we define $\alpha'_t(j)$ as

$$\alpha'_t(j) = \frac{\tilde{\alpha}_t(j)}{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t-1})}. \tag{15}$$

As a consequence from equation (14),

$$\alpha_t(j) = \frac{\alpha'_t(j) P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t-1})}{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t})}. \tag{16}$$

Then from equations (9) and (10), we obtain the recursion

$$\alpha'_{t+1}(k) \quad = \quad \begin{cases} \sum_j p_{jk} \alpha_t(j) f_{\boldsymbol{\theta}_k}(y_{t+1}) & \text{if } t+1 \in \boldsymbol{O}bs \\ \sum_j p_{jk} \alpha_t(j) & \text{if } t+1 \in \boldsymbol{M}is. \end{cases}$$

From equation (15),

$$\sum_k \alpha'_{t+1}(k) = \frac{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t+1})}{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t})}, \tag{17}$$

and from equation (16), $\alpha_{t+1}(k)$ is computed as

$$\alpha_{t+1}(k) = \frac{\alpha'_{t+1}(k)}{\sum_l \alpha'_{t+1}(l)}.$$

As a consequence from equation (17) applied recursively, the likelihood is given by

$$P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs}) = \prod_t \sum_j \alpha'_t(j) \tag{18}$$

The quantity $\beta_t(j)$ in the *backward* recursion is related to $\tilde{\beta}_t(j)$ as follows:

$$\beta_t(j) = \frac{\tilde{\beta}_t(j)}{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, t < u \leq n} | \{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t})}. \tag{19}$$

For each state k and each time $t < n$, we define $\beta'_t(j)$ as

$$\beta'_t(j) = \frac{\beta_t(j) P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t+1})}{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t})}. \tag{20}$$

Consequently, from equations (19) and (20),

$$\beta'_t(j) = \frac{\tilde{\beta}_t(j) P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs, u \leq t+1})}{P(\{Y_u = y_u\}_{u \in \boldsymbol{O}bs})}. \tag{21}$$

Then from equations (12) and (13), it leads to the recursion

$$\beta'_t(j) = \begin{cases} \sum_k p_{jk} \beta_{t+1}(k) f_{\boldsymbol{\theta}_k}(y_{t+1}) & \text{if } t+1 \in \boldsymbol{O}bs \\ \sum_k p_{jk} \beta_{t+1}(k) & \text{if } t+1 \in \boldsymbol{M}is. \end{cases}$$

Using equations (17) and (20), $\beta_t(j)$ is computed as

$$\beta_t(j) = \frac{\beta'_t(j)}{\sum_l \alpha'_{t+1}(l)}.$$

The quantities $P(S_t = j | \boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})$, denoted by $\xi_t(j)$, and $P(S_t = j, S_{t+1} = k | \boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})$, denoted by $\gamma_t(j,k)$, are derived as in Devijver (1985). From the classical equations

$$\xi_t(j) = \frac{\tilde{\alpha}_t(j) \tilde{\beta}_t(j)}{P(\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})}$$

and

$$\gamma_t(j,k) = \frac{\tilde{\alpha}_t(j) p_{jk} f_{\boldsymbol{\theta}_k}(y_{t+1}) \tilde{\beta}_{t+1}(k)}{P(\boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})},$$

we derive the following expression for the smoothed probabilities:

$$\begin{aligned} \xi_t(j) &= \alpha_t(j) \beta_t(j) \text{ and} \\ \gamma_t(j,k) &= \frac{\alpha_t(j) p_{jk} f_{\boldsymbol{\theta}_k}(y_{t+1}) \beta_{t+1}(k)}{\sum_l \alpha'_{t+1}(l)}. \end{aligned}$$

M step

The maximisation of the quantity $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(m-1)})$ with respect to $\boldsymbol{\lambda}$ leads to the standard update formula for $\boldsymbol{p}$

$$\hat{p}_{jk} = \frac{\sum_t P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j, S_{t+1} = k | \boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})}{\sum_t P_{\boldsymbol{\lambda}^{(m-1)}}(S_t = j | \boldsymbol{Y}_{\boldsymbol{O}bs} = \boldsymbol{y}_{\boldsymbol{O}bs})}.$$

The estimation formula for $\boldsymbol{\theta}_j$ depends on the family $\{f_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$. The formulas of Redner and Walker (1984) for independent mixtures in the exponential family of distributions

$$f_{\boldsymbol{\theta}}(y) = \frac{b(y)}{a(\boldsymbol{\theta})} e^{\boldsymbol{\theta}^t T(y)}$$

can be easily extended to the dependent case. The natural parameter $\mathbf{\Phi}_j = E_{\boldsymbol{\theta}}[T(Y)]$ is estimated by

$$\frac{\sum_{t \in \boldsymbol{Obs}} \xi_t(j) T(y_t) + [\sum_{t \in \boldsymbol{Mis}} \xi_t(j)] \mathbf{\Phi}_j^{(m-1)}}{\sum_t \xi_t(j)}.$$

For instance, for the Gaussian family with mean $\boldsymbol{\mu}$ and variance $\mathbf{\Sigma}$, this leads to the updating formulas

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{t \in \boldsymbol{Obs}} \xi_t(j) y_t + [\sum_{t \in \boldsymbol{Mis}} \xi_t(j)] \boldsymbol{\mu}_j^{(m-1)}}{\sum_t \xi_t(j)}$$

$$\hat{\mathbf{\Sigma}}_j = \frac{\sum_{t \in \boldsymbol{Obs}} \xi_t(j)(y_t - \hat{\boldsymbol{\mu}}_j)(y_t - \hat{\boldsymbol{\mu}}_j)^t + [\sum_{t \in \boldsymbol{Mis}} \xi_t(j)][\Sigma_j^{(m-1)} + (\boldsymbol{\mu}_j^{(m-1)} - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{\mu}_j^{(m-1)} - \hat{\boldsymbol{\mu}}_j)^t]}{\sum_t \xi_t(j)}.$$

References

[1] H. Akaike (1973). Information theory as an extension of the maximum likelihood theory. In B.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest.

[2] L.E. Baum, T. Petrie, G. Soules, and N. Weiss (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

[3] J.M. Bernardo and A.F.M. Smith (1994). *Bayesian Theory*. Chichester: Wiley.

[4] C. Biernacki, G. Celeux, and G. Govaert (2001). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

[5] C. Biernacki, G. Celeux, G. Govaert, F. Langrognet, G. Noulin, and Y. Vernaz (2003). Mixmod : un logiciel pour les modèles de mélange en classification et en analyse discriminante. In *XXXVèmes Journées de Statistique. Lyon*.

[6] S. Boucheron and E. Gassiat (2005). *Inference in hidden Markov models*, chapter order estimation. Springer-Verlag. edited by O. Cappé, E. Moulines and T. Rydén.

[7] G. Celeux and J. Clairambault (1992). Estimation de chaînes de Markov cachées : méthodes et problèmes. In *Actes des journées thématiques Approches markoviennes en signal et images. GDR signal-images CNRS*, pages 5–20.

[8] G.A. Churchill (1989). Stochastic Models for Heterogeneous DNA Sequences. *Bulletin of Mathematical Biology*, 51:79–94.

[9] J. Clairambault, L. Curzi-Dascalova, F. Kauffmann, C. Médigue, and C. Leffler (1992). Heart rate variability in normal sleeping full-term and preterm neonates. *Early Human Development*, 28:169–183.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38.

[11] P. A. Devijver (1985). Baum's forward-backward Algorithm Revisited. *Pattern Recognition Letters*, 3:369–373.

[12] J.-B. Durand (2003). *Modèles à structure cachée : inférence, sélection de modèles et applications*. PhD thesis, Université Grenoble 1 - Joseph Fourier.

[13] Y. Ephraim and N. Merhav (2002). Hidden Markov processes. *IEEE Transactions on Information Theory*, 48:1518–1569.

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

[15] E. Gassiat (2002). Likelihood ratio inequalities with application to various mixtures. *Annales de l'Institut Henri Poincaré*, 38:897–906.

[16] E. Gassiat and C. Kéribin (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM P & S*, 4:25–52.

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

[18] C. Kéribin (2000). Consistent estimation of the order of mixture models. *Sankhya Series*, A 62:49–66.

[19] G.J. McLachlan and D. Peel (1997). On a resampling approach to choosing the number of components in normal mixture models. In L. Billard and N.I. Fisher, editors, *Computing Science and Statistics*, volume 28, pages 260–266. Fairfax Station, Virginia: Interface Foundation of North America.

[20] G.J. McLachlan and D. Peel (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley and Sons.

[21] L.R. Rabiner (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, February.

[22] R.A. Redner and H.F. Walker (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239.

[23] B.D. Ripley (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

[24] C.P. Robert, G. Celeux, and J. Diebolt (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics and Probability Letters*, 16(1):77–83.

[25] A. W. Robertson, S. Kirshner, and P. Smyth (2004). Downscaling of daily rainfall occurence over Northeast Brazil using a hidden markov model. *Journal of Climate*, 17(7):4407–4424.

[26] K. Roeder and L. Wasserman (1997). Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association*, 92(439):894–902.

[27] G. Schwarz (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

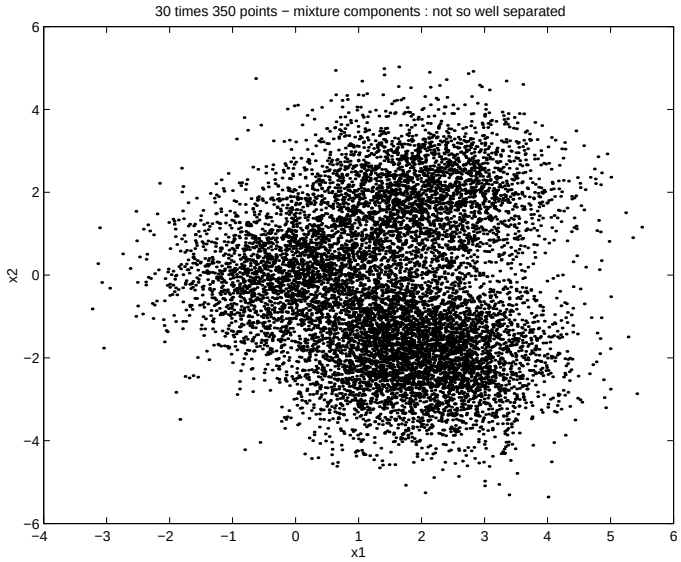
[28] P. Smyth (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72.

[29] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde (2000). Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistical Society (Series B)*, 64(4):583–639.

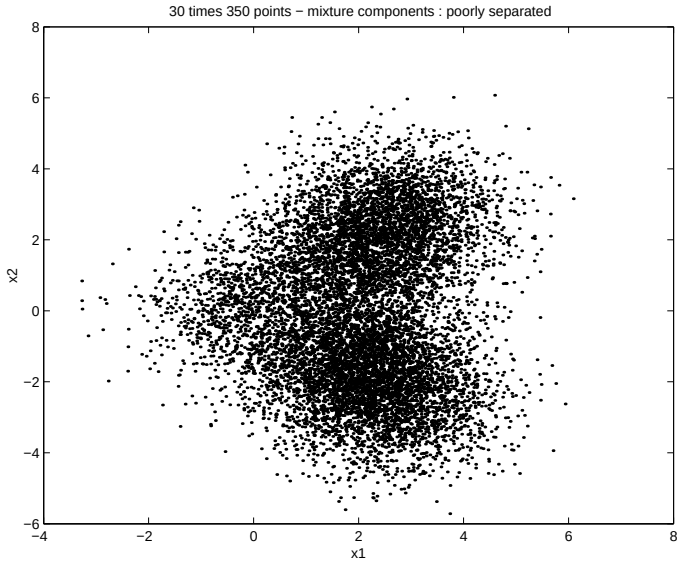
[30] Y. Yang (2005). Can the strengths of AIC and BIC be shared? A confict between model identification and regression estimation. *Biometrika*, 92:937–950.

[31] N. R. Zhang and D. O. Siegmund (2006). A modified Bayes information criterion with applications to the analysis of comparartive genomic hybridization data. *Biometrics*. to appear.

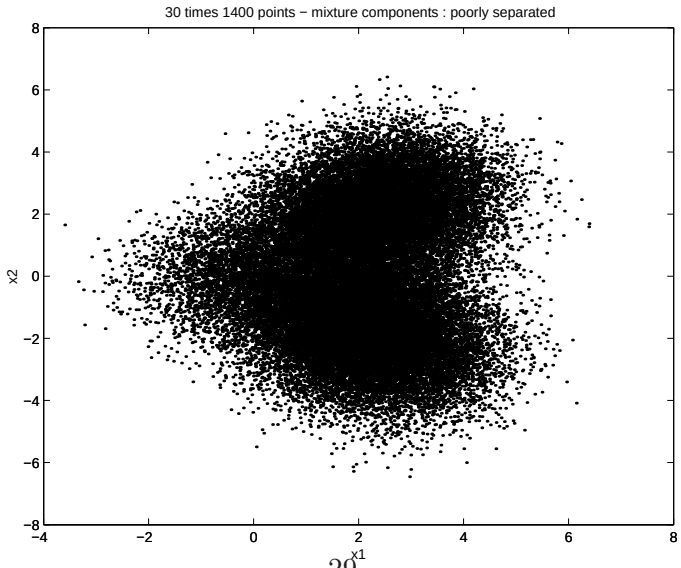
[32] P. Zhang (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 21(1):299–313.



(a)



(b)



(c)

Figure 2: *(a) Simulated data from 30 HMMs of length 350 with $K = 3$ hidden states; (b) Simulated data from 30 HMMs of length 350 with $K = 5$ hidden states; (c) Simulated data from 30 HMMs of length 1,400 with $K = 5$ hidden states.*

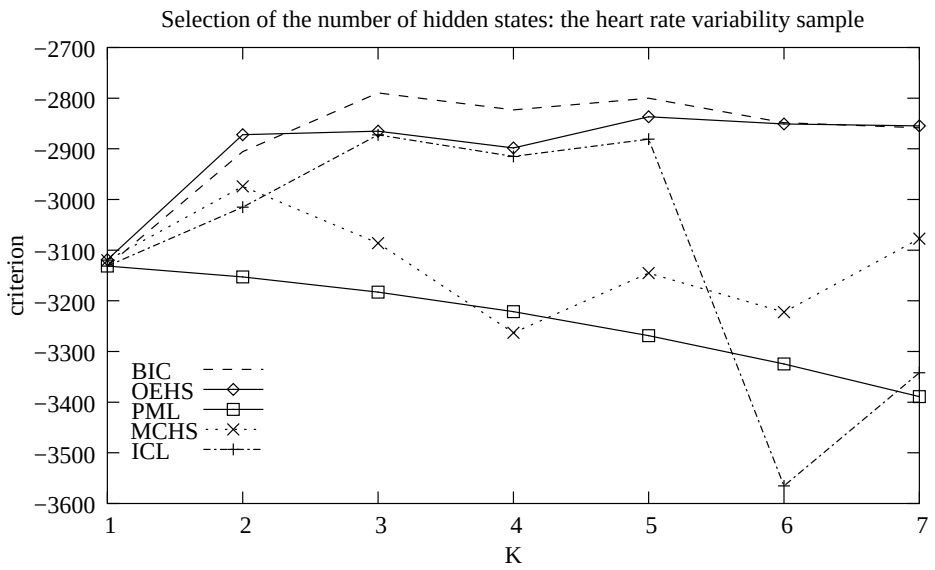


Figure 3: *Model selection for the heart rate variability sample. The values of the criteria OEHS, MCHS, BIC, PML and ICL are represented as a function of the number of hidden states K.*

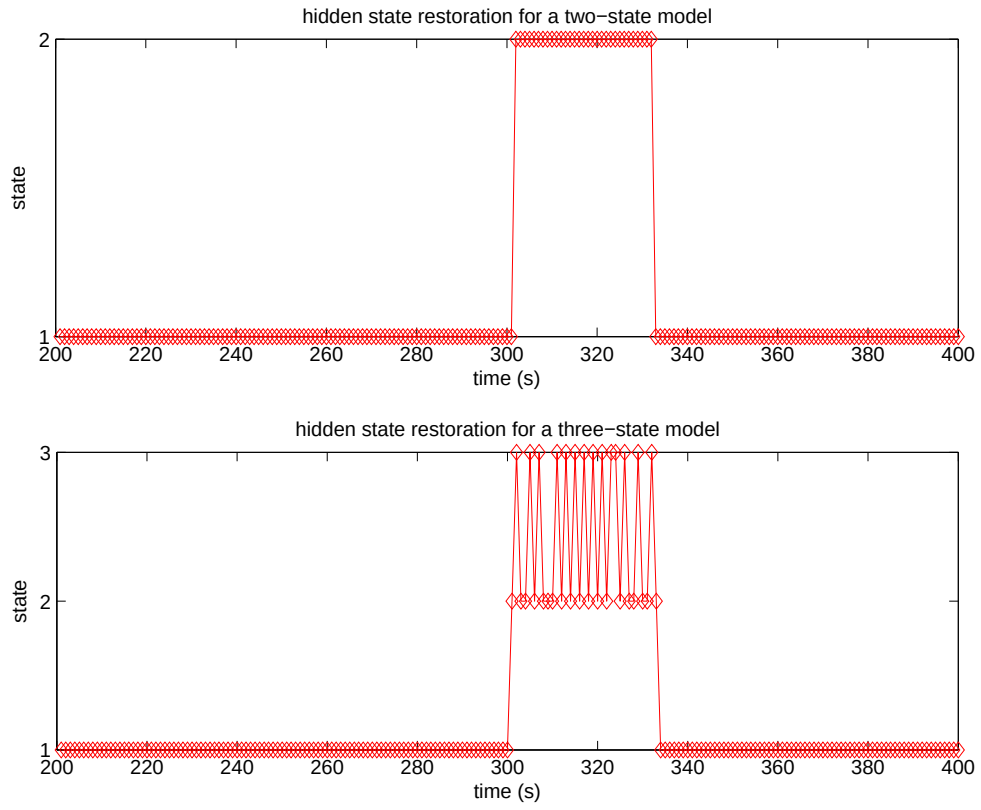


Figure 4: *Hidden state restoration for the heart rate variability dataset with the Viterbi algorithm (applied to a 200-point-wide window). Top: two-state HMM model. Bottom: three-state HMM model.*