Journal de la Société Française de Statistique
Vol. 151 No. 1 (2010)

Inference in Mixed Hidden Markov Models and Applications to Medical Studies

Titre: Inférence dans les modèles de Markov cachés à effets mixtes et applications médicales

Maud Delattre [1]

Abstract: The aim of the present paper is to document the need for adapting the definition of hidden Markov models (HMM) to population studies, as well as for corresponding learning methodologies. In this article, mixed hidden Markov models (MHMM) are introduced through a brief state of the art on hidden Markov models and related applications, especially focusing on disease related problems. Making the main assumption that a given pathology can be considered at different stages, hidden Markov models have for example already been used to study epileptic activity or migraine.

Mixed-effects hidden Markov models have been newly introduced in the statistical literature. The notion of mixed hidden Markov models is particularly relevant for modeling medical symptoms, but the data complexity generally requires specific care and the available methodology for MHMM is relatively poor. Our new approach can be briefly described as follows. First, we suggest to estimate the population parameters with the SAEM (Stochastic Approximation EM) algorithm, which has the property to converge quickly. The well-known forward recursions developed for HMM allow to compute easily the complete likelihood at each step of the MCMC procedure used within SAEM. Then, for dealing with the individuals, we suggest to estimate each set of individual parameters with the MAP (Maximum A Posteriori) of the parameter distributions. Finally, the hidden state sequences are decoded using the Viterbi algorithm. Some Monte-Carlo experiments are presented to illustrate the accuracy of our algorithms.

Résumé : Cet article veut montrer la nécessité d'étendre la définition des modèles de Markov cachés (HMM), ainsi que leurs méthodes d'estimation au cadre des études de population. Nous motivons alors les intérêts des modèles de Markov cachés à effets mixtes (MHMM) au travers d'un état de l'art succinct sur les modèles de Markov cachés et leurs nombreuses applications. Nous nous limiterons à des problématiques médicales. Dans ce cadre, les modèles de Markov cachés supposent que l'évolution des maladies peut s'interpréter à travers différents états. En effet, la distinction de divers stades évolutifs dans la maladie justifie l'application de modèles de Markov cachés à certaines pathologies, comme cela a déjà été le cas pour la migraine, ou encore l'épilepsie.

La définition des modèles de Markov cachés à effets mixtes est très récente. Ces nouveaux modèles sont des candidats intéressants pour la modélisation de symptômes. Les données utilisées sont complexes par leur structure et nécessitent toutefois une démarche d'analyse particulière. En outre, les méthodes d'apprentissage pour les MHMM restent peu nombreuses. Notre démarche est la suivante. Pour commencer, nous proposons d'estimer les paramètres de population au moyen de l'algorithme SAEM (Stochastic Approximation EM), dont la convergence est rapide. La procédure forward développée pour les HMM permet ici un calcul simple de la vraisemblance complète à chaque étape de la procédure MCMC de l'algorithme. Ensuite, les paramètres individuels sont obtenus par maximisation a posteriori de leur distribution. Enfin, les séquences d'états les plus probables pour chaque individu sont estimées par l'algorithme de Viterbi. Une étude par simulations Monte-Carlo illustre les propriétés de nos algorithmes.

Keywords: hidden Markov models, mixed-effects, longitudinal data, stochastic approximation EM, forward recursions, maximum a posteriori, Viterbi algorithm
Mots-clés : modèles de Markov cachés, effets mixtes, données longitudinales, algorithme SAEM, algorithme forward, maximum a posteriori, algorithme de Viterbi
AMS 2000 subject classifications: 62-02, 62F10

[1] Département de Mathématiques, Faculté des Sciences d'Orsay, Université Paris-Sud 11, F-91405 Orsay Cedex.
E-mail: maud.delattre@math.u-psud.fr

 ISSN: 2102-6238

1. Hidden Markov models

1.1. Definition

Hidden Markov models first appeared in the statistical literature in the late 60's, in a series of articles from Baum and coworkers ([7, 6, 8]). Those new models weren't referred to as "hidden Markov models" yet. The expression "probabilistic functions of finite state Markov chains" was rather used, reflecting quite well the definition of hidden Markov models. HMM applications, especially to speech recognition, abounded in the 70's and were at the origin of some methodological developments for learning these new models. In 1989, Rabiner published a tutorial ([16]) in which hidden Markov models were rigorously defined. This paper also clarified the related methodologies and gave several application examples devoted to speech recognition problems.

In hidden Markov models, a double process is assumed, among which only one is observed. Consequently, two levels are separately considered in the definition of such models. First comes the definition of the hidden process, which is a Markov process, generally assumed to have memory one; second comes the definition of the observations' process conditional on the hidden one. In this section, we will consider a parametric framework and Ψ will denote the vector of all HMM parameters.

Let $\{Z_j\}_{j\in\mathbb{N}\star}$ be a stationary Markov chain with a discrete and finite state space $\mathbf{S} = \{1,\ldots,S\}$. In the following, let π be the stationary distribution of the chain and let A denote the $S \times S$ transition matrix of the Markov process, and $\forall s, s' \in \mathbf{S}^2$, let also $a_{s,s'}$ be the transition probability associated with the transition from state s to state s':

$$a_{s,s'} = \mathbb{P}_\Psi\left(Z_{j+1} = s' | Z_j = s\right); \forall j \geq 1$$

where

$$\sum_{s'=1}^{S} a_{s,s'} = 1; \forall s \in \mathbf{S}$$

In hidden Markov models, a second process $\{Y_j\}_{j\in\mathbb{N}\star}$ enables inference on the latent process. More precisely, the jth observation is assumed to be a probabilistic function of the jth hidden visited state.

For discrete observations, emission probabilities are introduced to specify how the observations are related to the state sequence. Let $\mathbf{O}$ be the observation space, and let $b_{o,s}$ be the probability associated to observation $o \in \mathbf{O}$ while current (unknown) state is $s \in \mathbf{S}$:

$$b_{o,s} = \mathbb{P}_\Psi\left(Y_j = o | Z_j = s\right); \forall j \in \mathbb{N}\star$$

The emission probabilities are such that:

$$\sum_{o\in\mathbf{O}} b_{o,s} = 1; \forall s \in \mathbf{S}$$

For example, when the observations are count data, Poisson distributions with parameters λ_1, λ_2, $\ldots$, λ_S are usually chosen to model emissions in states 1, 2, $\ldots$, S respectively. The model for observations would thus read:

$$b_{o,s} = \mathbb{P}_{\lambda_s}\left(Y_j = o | Z_j = s\right) = \exp(-\lambda_s)\frac{\lambda_s^o}{o!}; \forall s \in \mathbf{S}$$

 ISSN: 2102-6238

An additional assumption holds: conditionally to Z_j, Y_j is independent of $Y_1,\ldots,Y_{j-1},Y_{j+1},\ldots$ and $Z_1,\ldots,Z_{j-1},Z_{j+1},\ldots$.

From the above definitions and assumptions, the general expression of the likelihood in HMMs becomes:

$$L(y_1,\ldots,y_J;\Psi)=\sum_{z_1,\ldots,z_J\in S^J}\pi(z_1)\prod_{j=1}^{J}b_{y_j,z_j}\prod_{j=1}^{J-1}a_{z_j,z_{j+1}} \tag{1}$$

where J is the length of the observed sequence.

Remark: extension to continuous observations is straightforward by defining conditional emission densities, for example a Gaussian distribution with parameters $(m_s;\sigma_s^2)$ in state $s\in\mathbf{S}$.

1.2. Learning methods

Infering hidden Markov models is challenging, mostly due to the complex expression of the likelihood and to the non observable visited states. As a consequence, hidden Markov models are associated with three "basic problems". Quoting from [16], (i) computing the likelihood, (ii) estimating the model parameters (emission probabilities, transition probabilities, and possibly the probability distribution of the states at time 1), and (iii) decoding the most probable state sequence for a given sequence of observations. Some algorithms are referenced and discussed in [16], giving potential solutions to (i), (ii) & (iii). Among those presented in [16], the forward procedure, the Baum-Welch algorithm and the Viterbi algorithm are the most relevant ones, applying respectively to (i), (ii) & (iii). The Baum-Welch algorithm is an EM-type algorithm which is expected to compute the maximum likelihood estimator (MLE). The question of the consistency of the MLE has been largely investigated. This is however a complex problem, and very strong assumptions were usually required to get the MLE consistency. Among other works, those of Leroux ([15]) and Douc and Matias ([11]) could be cited. More recently, Douc, Moulines, Olsson and Van Handel demonstrated the consistency of the MLE under very weak assumptions in [12]. Their result even holds in situations where the state space is not compact.

2. From hidden Markov models to mixed hidden Markov models

Many authors suggested applying hidden Markov models to deal with some biological problems. Hidden Markov models have become a very successful modeling tool in molecular biology, and applications to genetics abound in the statistical literature. Hidden Markov models have also been used in epidemiology, as an alternative to compartimental SIR (*Susceptible-Infectious-Recover*) and SIS (*Susceptible-Infectious-Susceptible*) models, to study the propagation of epidemics or infections in specific populations ([17, 9]). In this context, the observations consist of counts of infected patients at successive time points. As such phenomenons are characterized by patient to patient transmission, successive outcomes can't be considered as independent outcomes and assuming a Markov structure is of strong interest. Sometimes, the studied disease is asymptomatic on some patients. Hidden Markov models could thus be an adapted way to describe the infection latent process, through the estimation of important epidemiological parameters, or even through the distinction of periods with different transmission rates. For example, Le Strat and Carrat were interested in influenza-like illness (I.L.I.) and poliomyelitis in 1999 ([17]). In [17], I.L.I. incidence

 ISSN: 2102-6238

rates were modeled as a two-state Gaussian hidden Markov model, leading to a clear distinction between epidemic and non epidemic periods. The number of monthly Poliomyelitis cases were analyzed in the same manner, with a two-state Poisson hidden Markov model. In a same way, Cooper and Lipsitch were interested in nosocomial infections ([9]) and proved that models with a latent Markov process granted the best fit of their data. Then, some authors showed that hidden Markov models could be particularly relevant for the study of chronic illnesses, such as migraine ([5]), multiple sclerosis ([2, 4]) or epilepsy ([1]). Here, the transitions between some unobserved states, whom total number is possibly predefined, are supposed to describe the evolution of disease symptoms, like the daily numbers of seizure counts in the context of epilepsy ([1]), the monthly numbers of lesion counts in the context of multiple sclerosis ([2, 4]), or headache scores when dealing with migraine ([5]).

To study diseases, the interests of hidden Markov models are numerous. First, those particular models are quite easily interpretable, and appear to show up similarities with the biological process that governs the pathologies. The Markov states are thus associated with distinct stages or seriousness degrees for the studied illness, and the assumption is that patients alternate periods in those stages. As an example, patients with multiple sclerosis seem to undergo relapsing and remitting periods ([2, 4]). In the same way, Albert ([1]) assumes the epileptic patients to go through two distinct stages, namely a low and a high seizure susceptibility. In related clinical trials, markers are used to support disease diagnosis, and explaining their value by indirectly observed illness stages seems to have a biological meaning; at least this approach is widely used. Seeing this, the use of latent states, typically through mixture models, to study disease dynamics on specific patients is natural. It is also quite reasonable to assume that consecutive values of a biomarker for a given patient are interdependent. For example, past events or repeated past passages to acute forms of the disease could reinforce susceptibility to the illness. It is thus justified to enrich the mixture with a first-order memory, leading to hidden Markov models. Having a well-founded biological interpretation, we could also imagine hidden Markov models could improve the understanding of the process underlying some more obscure pathologies.

Therefore, hidden Markov models have conceptual validity in some disease studies. The estimated model parameters help thus to interpret the disease process at several levels. First, through the emission distributions, they give some idea of the way the biomarker values are related to the hidden states. Second, the estimated transition probabilities help to see how state changes are frequent in the studied population. Interpretation could even be carried on by including covariates and regression variables in the parametrization of the model.

However, modeling disease using hidden Markov models is not straightforward. In particular, while hidden, the "design" of the underlying Markov process could be challenging. When enough knowledge on the disease of interest exists, the number of hidden states can be a priori fixed; but most often, the number of hidden illness stages is unknown. Several numbers of states have to be tried and adapted selection criteria are required. As an example, hidden Markov models from two to six states are tried to model I.L.I. data in [17], and the BIC criteria is chosen to discriminate the most adapted model from the others, leading to a five-state Gaussian hidden Markov model. Le Strat stresses the lack of interpretability of such a result.

More specific difficulties occur when modeling the outcomes of clinical studies, mainly due to their structure. Indeed, several patients are included, and are subject to repeated measurements.

As hidden Markov models are a possible way to analyze one particular sequence of data, the

 ISSN: 2102-6238

first approach consisted in considering as many hidden Markov models as included patients. Each individual set of parameters was therefore estimated independently of the others. Albert followed this approach to epileptics'seizure count data ([1]) and to multiple sclerosis data ([2]). However, by continuing Albert's work on multiple sclerosis, Altman ([4]) underlined estimation inaccuracy, and noted the obtained estimates were always associated with large standard errors.

Clearly, the individual fit approach to longitudinal data has the major drawback of incorrectly capturing the heterogeneity among patients. Indeed, the complete set of individual estimates only gives a limited summary of the variation or heterogeneity of the individual parameters. The need for "... a model [that would] describe all patient's data simultaneously" was therefore argued for the first time in Altman's article ([3]). On the same idea as mixed models, the heterogeneity characterizing the data would be finely taken into account by including i.i.d. random effects in each patient's hidden Markov model parameters definition. It would also be a way to foresee possible correlations between parameters. This way, Altman supposed defining a hidden Markov model with random parameters would help to increase the precision of the estimates, and would best capture the potential variation among patients. Those remarks, dating back to 2005, are at the origin of mixed hidden Markov models.

3. Mixed hidden Markov models

3.1. Definition

A rigorous definition of mixed hidden Markov models by Altman followed in 2007 in [3]. Parallel to this work, Ip and coworkers also published an article on mixed hidden Markov models in 2007 ([13]). In both papers, mixed hidden Markov models appear as an extension of "classical" hidden Markov models to deal with the specific contexts using a population approach.

Mixed hidden Markov models include several levels of definition. Assume we have at our disposal data from n subjects. A hidden Markov model is used for each individual set of data, while the parameters for each individual model are assumed to be random with a common probability distribution. As for the definition of HMM, we will consider a parametric framework. Using the same notations, $\mathbf{O}$ is the common observation space, and $\mathbf{S}$ is the common state space.

3.1.1. Definition of n "distinct" hidden Markov models

The first step of a MHMM's definition consists in specifying a hidden Markov model for the observations of each of the n subjects. More precisely, the distribution of the observations for each individual is based on a Markov chain, which sequence of visited states is unknown. Let us restrict to subject i ($1 \leq i \leq n$). Let n_i be the number of observations for this subject, and let $\mathbf{Y}_i = (y_{i1} \dots y_{i,n_i})^T$ and $\mathbf{Z}_i = (z_{i1} \dots z_{i,n_i})^T$ be respectively the sequence of observations for individual i and his sequence of hidden states. As MHMM definition mainly goes through the specification of individual hidden Markov models, as many sets of parameters as subjects are required instead of only one set of parameters Ψ for HMM. Let Ψ_i denote the vector of parameters for subject i. Typically, Ψ_i is part of the definition of

1. the emission distributions, via a series of emission probabilities for discrete observations:

$$b_{o,s}^{(i)} = \mathbb{P}_{\Psi_i}(y_{ij} = o | z_{ij} = s); \forall 1 \leq j \leq n_i; \forall o \in \mathbf{O}, \forall s \in \mathbf{S}$$

 ISSN: 2102-6238

2. the transition matrix

$$a_{s,s'}^{(i)} = \mathbb{P}_{\Psi_i}\left(z_{i,j+1} = s' | z_{ij} = s\right); \forall j \geq 1, \forall (s,s') \in \mathbf{S}^2$$

3.1.2. *Model for the individual parameters*

The n vector of individual parameters Ψ_i have a same probability distribution. The parameters θ of this population distribution are the so-called population parameters. We will consider a linear Gaussian model for the (transformed) individual parameters that can include covariates:

$$\begin{cases} h(\Psi_i) = \mu + C_i\beta + D_i\eta_i \\ \eta_i \underset{i.i.d.}{\backsim} \mathcal{N}(0,\Omega) \end{cases}$$

where h is a vector of link functions, C_i and D_i are known matrices of covariates for individual i, μ and β are unknown vectors of fixed effects, and Ω captures the variability of individual behaviors that the covariates can't explain themselves. Here, $\theta = (\mu, \beta, \Omega)$.

With such a hierarchical definition, a single statistical model describes the whole individuals' data simultaneously while taking into account the potential heterogeneity among patients.

The observed likelihood is given by

$$\begin{aligned} L(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n; \theta) &= \prod_{i=1}^{n} \int L(\mathbf{Y}_i, \Psi_i; \theta)\, d\Psi_i \\ &= \prod_{i=1}^{n} \int L(\mathbf{Y}_i | \Psi_i)\, L(\Psi_i; \theta)\, d\Psi_i \end{aligned} \tag{2}$$

where $L(\mathbf{Y}_i|\Psi_i)$ has a similar expression as the observed likelihood of a "classical" HMM given in (1). This observed likelihood cannot be computed in a closed form and this complex expression makes the model inference directly intractable.

3.2. *Inference in MHMM*

Inference in mixed hidden Markov models is not straightforward, and only some partial methods have already been suggested. Here, we will finally put forward a different way to grasp MHMM's learning.

3.2.1. *First steps*

Mixed hidden Markov models are somehow "new" models. As a consequence, their usage is documented in a very limited way and the related methodologies are not well established. The maximum likelihood approach could be used for estimating the population parameters but the complex expression for the likelihood makes its maximum difficult to locate. Mixed hidden Markov models can be viewed as missing data models where the visited states and the individual parameters are the non observed data. As a consequence, the EM algorithm seems to be a natural parameter estimation method for such models ([13]) but the E-step cannot be performed in a

 ISSN: 2102-6238

closed-form. Altman suggested alternative methods, such as quasi-Newton methods or Gaussian quadrature methods, or even the MCEM algorithm ([3]). Nevertheless, these algorithms are time expansive. Several days may be required to estimate the model parameters when the number of random effects in the model exceeds three. This forces her to restrict her attention to models involving random effects on the emission distribution only. Considering the problem of predicting mastisis prevalence in cows, Detilleux suggested to estimate the model parameters of a mixed hidden Markov model using a Gibbs sampler ([10]).

3.2.2. Our methodology

Performing parameter inference in mixed hidden Markov models has been underlined to be a complex problem. Knowledge of the population parameters is necessary to grasp the mean tendency, as well as its variability among individuals; but it is not enough when focusing on a particular individual. Using the population parameters alone could bias individual diagnosis. That's why it is important to divide the mixed hidden Markov models' problem into three main questions:

1. First naturally comes the question of estimating the population parameters.
2. Then the individual sets of parameters have to be estimated.
3. Estimating the most probable individual state sequences is the final issue to address.

In this paragraph, statistical methods dealing with the three points above are suggested. The use of the SAEM algorithm for estimating the population parameters is the most original part of our methodology, and is thus more detailed.

Population parameters' estimation In models such as mixed hidden Markov models, the E-step of the EM algorithm is not directly tractable. Then, we propose to adapt the MCMC-SAEM algorithm ([14]) to the mixed hidden Markov model setting. Each iteration of the algorithm can be decomposed into three steps. The non observed data are simulated (simulation step). These simulated data are used in a second step together with the observations to approximate the complete likelihood (stochastic approximation step). This likelihood is then be maximized to update the estimation of the parameters (maximization step).

In the context of mixed hidden Markov models, the first idea would be to consider the individual parameters (Ψ_i) and the Markov chains $(\mathbf{Z}_i)$ as the non observed data. Indeed, the conditional distribution of $(\Psi_i, \mathbf{Z}_i)$ can easily be simulated by MCMC and the complete likelihood $L(\mathbf{Y}_1, \ldots, \mathbf{Y}_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_n, \Psi_1, \ldots, \Psi_n; \theta)$ can easily be maximized.

Even if this first version of the algorithm can be implemented and gives good results, considering the Markov chain as a nuisance parameter of the model allows to propose a much more simple and efficient procedure. Note that a quick computation of the n individual likelihoods $L(\mathbf{Y}_i, \Psi_i; \theta)$ is the key of the algorithm. Indeed, the following decomposition allows many simplifications extremely useful for implementing the SAEM algorithm:

$$L(\mathbf{Y}_i, \Psi_i; \theta) = L(\mathbf{Y}_i|\Psi_i) L(\Psi_i; \theta) \tag{3}$$

Computing $L(\mathbf{Y}_i|\Psi_i)$ turns out to be easy by making use of the forward recursions that are part of the well-known Baum-Welch algorithm which allows computing the observed likelihood in

 ISSN: 2102-6238

hidden Markov models. Then, $L(\Psi_i;\theta)$ derive from the Gaussian distribution and is easy to compute and to maximize.

Let us describe iteration k of the algorithm. Here, θ_k denotes the current estimate of the population parameters.

1. *Simulation*
   The kth iteration begins with drawing $\Psi_i^{(k)}$ from the conditional distribution $p(\Psi_i|\mathbf{Y}_i;\theta_k)$ for all $1 \leq i \leq n$. The Hasting-Metropolis algorithm used for this simulation step requires to compute $L(\mathbf{Y}_i|\Psi_i;\theta_k)$ in a closed form for evaluating each acceptance probabilities. As mentioned above, computing this conditional likelihood is straightforward thanks to the forward procedure.

2. *Stochastic approximation*
   Follows a stochastic approximation of the log likelihood:

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\sum_{i=1}^{n} \log L\left(\mathbf{Y}_i, \Psi_i^{(k)}; \theta\right) - Q_{k-1}(\theta) \right]$$

   where $(\gamma_k)_{k\geq 0}$ is decreasing to 0 over iterations.
   $Q_k(\theta)$ can be written as the sum of two terms among which only one depends on parameter θ:

$$Q_k(\theta) = R_k + T_k(\theta)$$

   where

$$R_k = R_{k-1} + \gamma_k \left[\sum_{i=1}^{n} \log L\left(\mathbf{Y}_i|\Psi_i^{(k)}\right) - R_{k-1} \right]$$

   and

$$T_k(\theta) = T_{k-1}(\theta) + \gamma_k \left[\sum_{i=1}^{n} \log L\left(\Psi_i^{(k)}; \theta\right) - T_{k-1}(\theta) \right]$$

   Then, it is equivalent to maximize $Q_k(\theta)$ or $T_k(\theta)$ with respect to θ and our stochastic approximation step would just reduce in computing $T_k(\theta)$.

3. *Maximization*
   kth iteration ends in maximizing T_k to update the estimation of θ:

$$\theta_k = \underset{\theta}{\operatorname{argmax}}\, T_k(\theta)$$

Iterations of this procedure are repeated until numerical convergence of the sequence (θ_k) to some estimate $\hat{\theta}$ is achieved.

Computing the standard errors (s.e.) of the estimated parameter $\hat{\theta}$ requires computing the Fisher Information Matrix (F.I.M.). We propose to estimate the F.I.M. using the stochastic approximation procedure described in [14] and based on the Louis formula.

 ISSN: 2102-6238

Individual parameters' estimation After estimating the population parameters with the SAEM algorithm, each individual parameter estimate Ψ_i can be calculated through the MAP (Maximum A Posteriori) method:

$$\hat{\Psi}_i = \underset{\Psi_i}{\operatorname{argmax}} \quad p\left(\Psi_i | \mathbf{Y}_i; \hat{\theta}\right)$$

Such maximization for each individual requires some optimization procedure.

Remark: An alternative would be to estimate the conditional mean $\mathrm{E}\left(\Psi_i | \mathbf{Y}_i; \hat{\theta}\right)$ with the MCMC procedure used within the SAEM algorithm.

Most likely state sequences' decoding Once the individual parameters (Ψ_i) are estimated, each individual model can be considered separately and the optimal individual state sequences can be decoded using the Viterbi algorithm:

$$\hat{\mathbf{Z}}_i = \underset{\mathbf{Z}_i}{\operatorname{argmax}} \quad p\left(\mathbf{Z}_i | \mathbf{Y}_i, \hat{\Psi}_i\right)$$

4. Application

4.1. The model

Our simulations were inspired by the quite numerous studies on epileptic activity. Similarly to the works cited above, we assumed the existence of a hidden Markov chain, which would condition the intensity of the seizures in epileptic patients. The common intuition is the following. The first and the second states would respectively be associated with a low and a high epileptic activity. Periods in both states would thus alternate in epileptic patients. As in [1], the emission distributions are chosen to be Poisson distributions. This means that conditional to the state the number of daily seizures for a given epileptic patient is assumed to follow a Poisson distribution. Let $\lambda_1^{(i)}$ and $\lambda_2^{(i)}$ be individual i's Poisson parameters in state 1 and in state 2, with $\lambda_1^{(i)} < \lambda_2^{(i)}$. Let also $p_{11}^{(i)}$ and $p_{21}^{(i)}$ be individual i's transition probabilities associated respectively with the transitions from state 1 to state 1 and from state 2 to state 1.

Our model is the following:

$$\begin{aligned}
\operatorname{logit}(p_{11}^{(i)}) &= \gamma_1 + \eta_{1i} && (4)\\
\operatorname{logit}(p_{21}^{(i)}) &= \gamma_2 + \eta_{2i} && (5)\\
\log(\lambda_1^{(i)}) &= \log(\lambda_1) + \eta_{3i} && (6)\\
\log(\alpha^{(i)}) &= \log(\alpha) + \eta_{4i} && (7)\\
\lambda_2^{(i)} &= \lambda_1^{(i)} + \alpha^{(i)} && (8)
\end{aligned}$$

The random effects are assumed to be independent and normally distributed:

$$\eta_i = (\eta_{1i}, \eta_{2i}, \eta_{3i}, \eta_{4i}) \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega)$$

θ corresponds here to the concatenation of the fixed effects (γ_1, γ_2, λ_1, α) and the elements of the variance-covariance matrix Ω.

 ISSN: 2102-6238

4.2. *A first numerical experiment*

One dataset with 200 individuals and 100 observations per subject were simulated using the following values for the fixed effects: $\gamma_1 = 1.4$, $\gamma_2 = -1.4$, $\lambda_1 = 0.8$, $\alpha = 2.3$. The random effects were simulated assuming a diagonal variance-covariance matrix Ω with the following diagonal elements: $\omega^2_{\gamma_1} = 0.1$, $\omega^2_{\gamma_2} = 0.1$, $\omega^2_{\lambda_1} = 0.2$ and $\omega^2_{\alpha} = 0.1$.

Table 1 displays the results of the SAEM algorithm used for estimating the population parameters and their standard errors. The true values of the population parameters $\theta^\star$, the initial values θ_0 and the estimates $\hat{\theta}$ are given together with their estimated standard errors (s.e.) and relative standard errors (r.s.e.).

Table 1 shows the SAEM estimates are similar to the true values. On this particular example, the relative estimation error is less than 15% on the whole, except for parameter $\omega^2_{\gamma_2}$ (36%). We also note that the (relative) standard errors for each parameter are low, which is very encouraging, except for variance parameters $\omega^2_{\gamma_1}$ and $\omega^2_{\gamma_2}$ (48% and 41% respectively).

TABLE 1. *Estimation of the population parameters: the true values, the initial values, the estimations, their standard errors and relative standard errors.*

	$\theta^\star$	θ_0	$\widehat{\theta}$	s.e.	r.s.e. (%)
γ_1	1.4	0.4	1.41	0.058	4
γ_2	-1.4	-0.4	-1.45	0.06	4
λ_1	0.8	2	0.779	0.03	4
α	2.3	0.5	2.25	0.062	3
$\omega^2_{\gamma_1}$	0.1	0.4	0.113	0.055	48
$\omega^2_{\gamma_2}$	0.1	0.4	0.136	0.056	41
$\omega^2_{\lambda_1}$	0.2	0.4	0.202	0.029	14
ω^2_{α}	0.1	0.4	0.115	0.015	13

Figure 1 shows the sequences of estimated parameters (θ_k). One clearly sees that SAEM converges in very few iterations to a neighborhood of the "true" value used for simulating the data, even with a poor initialization. Moreover, it took only $6'$ on a laptop for estimating both the population parameters and the Fisher information matrix with this dataset.

Then we have estimated the individual parameters $(\Psi_i; 1 \leq i \leq 200)$ by computing the MAP estimates for each subject.

Finally, each individual state sequence was estimated with the Viterbi algorithm. As dealing with simulated datasets, the "true" state sequences are known. Even if this information is omitted during the whole inference process, true and estimated states can be compared. Figure 2 presents the results obtained with three typical subjects. On each graph, the (simulated) observations (daily seizures) are represented as a function of time (number of days). The true unknown states are displayed in the left column. The second column depicts the raw data, i.e. the only information available in the practice for inference. The right column displays the estimated states. We can observe a very good agreement between the true and the decoded states.

 ISSN: 2102-6238

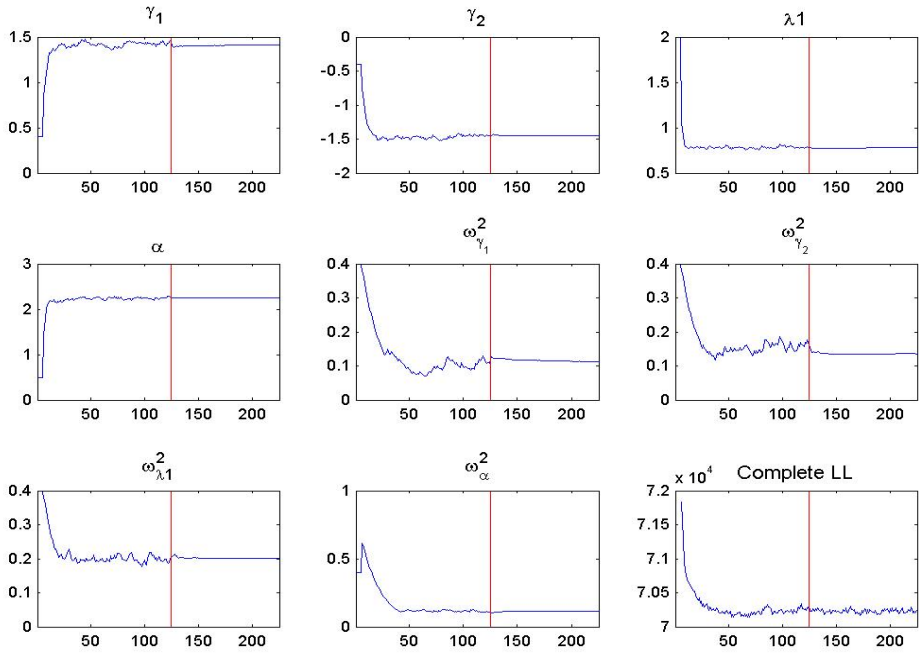

FIGURE 1. *Estimation of the population parameters: convergence of the SAEM algorithm.*

4.3. Monte Carlo study

This first result is encouraging but it was obtained on a particular dataset. Then a Monte Carlo study should confirm the good behavior of the proposed methodology. 100 datasets were simulated using the same design and the same population distribution. Then, the population parameters and their standard errors were estimated with SAEM for each dataset. For $m = 1, 2, \ldots, 100$, let $\hat{\theta}_m$ be the estimated vector of population parameters obtained with the mth simulated dataset and let $\widehat{\text{rse}}_m$ be their respective estimated standard-errors. For each model parameter, we have computed the mean estimated parameter $\overline{\theta}$, the mean estimated relative standard error $\overline{\text{rse}}$ and the relative standard deviation of the estimated parameters $\text{rsd}(\hat{\theta}\theta)$:

$$\overline{\theta} = \frac{1}{100}\sum_{m=1}^{100}\hat{\theta}_m \tag{9}$$

$$\overline{\text{rse}} = \frac{1}{100}\sum_{m=1}^{100}\widehat{\text{rse}}_m \tag{10}$$

$$\text{rsd}(\hat{\theta}) = 100 \times \sqrt{\frac{1}{100}\sum_{m=1}^{100}\left(\frac{\hat{\theta}_m - \theta^\star}{|\theta^\star|}\right)^2} \tag{11}$$

 ISSN: 2102-6238

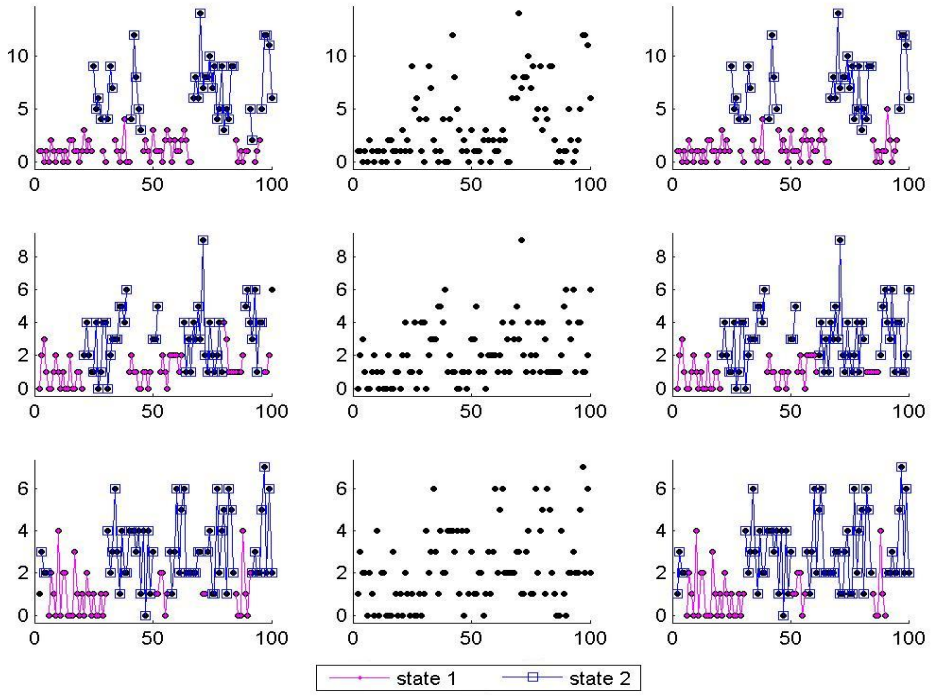

FIGURE 2. *State sequences of three typical subjects. Left: the observed data and the true unknown states, center: the observed data without the states, right: the estimated states.*

Here, $\theta^\star$ denotes the true values of the population parameters. Table 2 gives a brief summary of the Monte Carlo results.

Figures 3 and 4 display the empirical distributions of the 100 relative estimation errors defined as:

$$\text{REE}_m \quad = \quad 100 \times \frac{\hat{\theta}_m - \theta^\star}{\theta^\star} \tag{12}$$

Except for $\omega^2_{\gamma_1}$ and $\omega^2_{\gamma_2}$, one can observe that the estimates show very little bias and small ranges (table 2). The median REE for the fixed effects γ_1, γ_2, λ_1 and α remains between -5% and 5% (figures 3). The estimated variances ω_{λ_1} and ω_α are also very well estimated. The variances ω_{γ_1} and ω_{γ_2} are more difficult to estimate accurately. Indeed, the REE boxplots suggest quite important relative RMSE (root mean square errors) for those two parameters (44% and 51%) but this apparent estimation difficulty is in accordance with the estimated relative standard errors for those two parameters (54% and 58%).

More generally, one can remark the very good agreement between the estimated standard errors and the empirical standard deviations. The empirical standard deviations obtained from simulated

 ISSN: 2102-6238

TABLE 2. *Estimation of the population parameters: the true values, the means and the relative standard deviations of the estimated parameters, the mean estimated relative standard errors.*

	$\theta\star$	$\overline{\theta}$	rsd$(\hat{\theta})$ (%)	$\overline{\text{rse}}$ (%)
γ_1	1.4	1.394	4	4
γ_2	-1.4	-1.414	4	4
λ_1	0.8	0.778	3	4
α	2.3	2.292	3	3
$\omega^2_{\gamma_1}$	0.1	0.112	44	54
$\omega^2_{\gamma_2}$	0.1	0.121	51	58
$\omega^2_{\lambda_1}$	0.2	0.200	14	14
ω^2_{α}	0.1	0.098	14	13

data allow to evaluate the uncertainty of the estimated parameters. Of course, these empirical standard deviations cannot be computed in the practice when only one dataset is available and when the true population parameters are unknown. Nevertheless, one can have confidence with the estimated s.e. provided by the algorithm for evaluating the uncertainty of the estimated parameters.

These numerical results suggest that our algorithm produces unbiased and consistent population parameter estimates and standard errors in large databases. A theoretical study of the statistical property of the maximum likelihood population estimates is beyond the scope of this paper, but we will consider this issue in future works. On the other hand, more exhaustive studies should be led considering more difficult and more realistic contexts than the one here.

4.4. Technical remarks

The proposed methodology for MHMM has been implemented in the MONOLIX software (http://software.monolix.org). All the numerical examples were performed with MONOLIX 3.1.

In Monolix, it is possible to choose the number of Markov chains used for the SAEM algorithm. Here, two Markov chains were used instead of only one chain. That allowed to slightly improve the convergence of the algorithm by reducing its stochastic behavior.

Each initial guess for the SAEM algorithm was randomly chosen for each Monte Carlo run.

The initial probability distribution of the hidden Markov chain was not estimated. It was assumed that $\pi_1^{(i)} = \pi_2^{(i)} = \frac{1}{2}$.

5. Conclusion and perspectives

A brief state of the art on HMM applications to disease progression data shows that hidden Markov models are a reasonable modeling tool in this context. However, longitudinal data need models able to take account of the existing heterogeneity between individuals. This remark recently lead to the use of mixed hidden Markov models. However, related algorithms initially first tackled the population parameter estimation only including a small number of random effects. We suggested a new and complete inference methodology. The originality of our work consists in the use of the SAEM algorithm for estimating the model population parameters. A Monte Carlo study showed its good practical properties. More precisely, the SAEM algorithm converges to a neighborhood of the good parameter values in very few iterations even when the initial guess is poor. The

 ISSN: 2102-6238

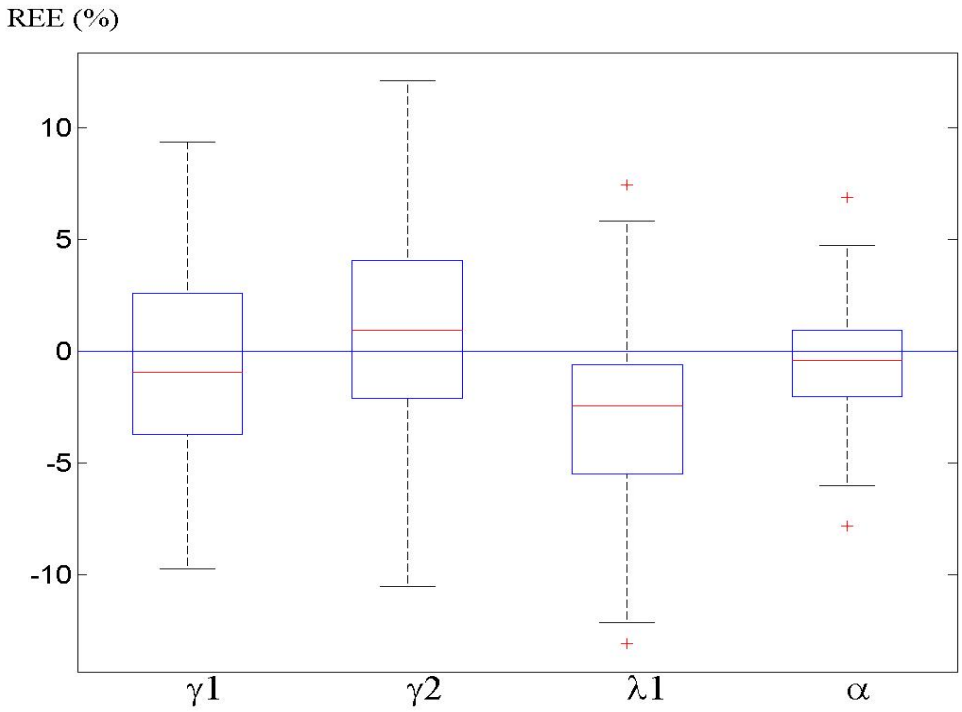

FIGURE 3. *Estimation of the fixed effects: empirical distribution of the relative errors of estimations.*

estimation process is fast, even with large datasets. The impact of the data size and the theoretical properties of the algorithm keep to be rigorously studied in future works. The main problems to tackle would be a precise analysis of the convergence of the algorithm as well as the statistical properties of the maximum likelihood estimate in MHMM.

From a practical point of view those new models seem to offer very promising statistical applications. More precisely, mixed-effects hidden Markov models could help for a more finely analysis of clinical trials, when the collected data often consist of longitudinal count data and when one suspects several hidden states. Having a measure of interest, the most popular approach consists of mean comparisons, between groups of patients or treatment periods. However, the classical comparison methods could sometimes lead to improper conclusions. Assume a finite set of hidden stages give a plausible interpretation for the dynamics of the studied pathology, then an inappropriate choice for the statistical model (i.e. a model ignoring the transitions between distinct stages) would not catch enough information on the phenomenon observed. We could imagine a treatment effect occurs at transition level and a variation of the time spent in one specific hidden state once entering this state could constitute the only difference between treatment groups or between treatment periods, without modifying the observations' distribution in any state. Then, a simple comparison between means of observed outcomes would either fail in bringing to light a significant treatment effect or show an overestimated or underestimated treatment induced change. Mixed hidden Markov models including the treatment group or the treatment dose as covariates in

 ISSN: 2102-6238

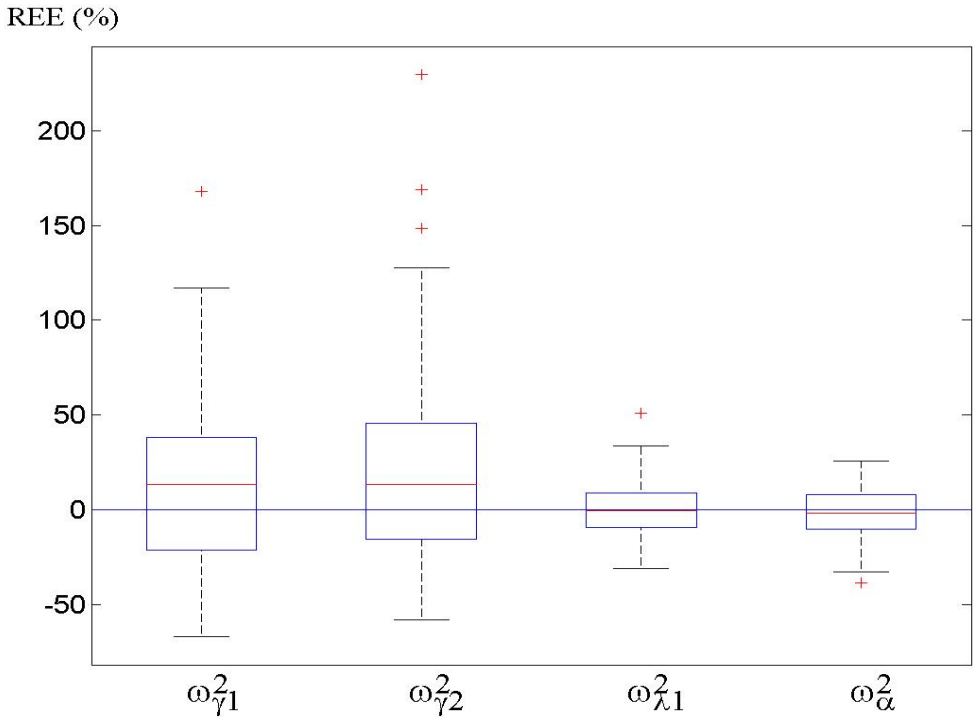

FIGURE 4. *Estimation of the variance parameters: empirical distribution of the relative errors of estimations.*

the definition of both the transition model and the observation model, would thus help to capture the true treatment effect and adequately locating it.

References

[1] Albert. A two state markov mixture model for a time series of epileptic seizure counts. *Biometrics*, 37:1371–1381, 1991.

[2] Albert, McFarland, Smith, and Frank. Time series for modelling counts from relapsing remitting disease : application to modelling disease activity in multiple sclerosis. *Statistics in Medicine*, 13:453–466, 1994.

[3] Altman. Mixed hidden markov models : an extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 2007.

[4] Altman and Petkau. Application of hidden markov models to multiple sclerosis lesion count. *Statistics in Medicine*, 24:2335–2344, 2005.

[5] Anisimov, Maas, Danhof, and Della Pasqua. Analysis of responses in migraine modelling using hidden markov models. *Statistics in Medicine*, 26:4163–4178, 2007.

[6] Baum and Eagon. An inequality with applications to statisical estimation for probabilistic functions of markov processes and to a model for ecology. 1966.

[7] Baum and Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 1966.

[8] Baum, Petrie, Soules, and Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41:164–171, 1970.

[9] Cooper and Lipsitch. The analysis of hospital infection data using hidden markov models. *Biostatistics*, 5:223–237, 2004.

 ISSN: 2102-6238

[10] Detilleux. The analysis of disease biomarker data using a mixed hidden markov model. *Genet. Sel. Evol.*, pages 491–509, 2008.

[11] Douc and Matias. Asymptotics of the maximum likelihood estimator for general hidden markov models. *Bernoulli*, 7:381–420, 2001.

[12] Douc, Moulines, Olsson, and Van Handel. Consistency of the maximum likelihood estimator for general hidden markov models. *The Annals of Statistics*. to appear.

[13] Ip, Snow Jones, Zhang, and Rijmen. Mixed effects hidden markov models. *Statistics in Medicine*, 2007.

[14] Kuhn and Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM : Probability and Statistics*, 8:115–131, 2004.

[15] Leroux. Maximum-likelihood estimation for hidden markov models. *Stochastic Processes and their Applications*, 40:127–143, 1992.

[16] Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[17] Le Strat and Carrat. Monitoring epidemiologic surveillance data using hidden markov models. *Statistics in Medicine*, 18:3463–3478, 1999.

 ISSN: 2102-6238