Bernoulli **11**(11), 2005, 103–129

Consistent and asymptotically normal parameter estimates for hidden Markov mixtures of Markov models

PIERRE VANDEKERKHOVE

Laboratoire d'Analyse et Mathématiques Appliquées, UMR 8050, University of Marne-la-Vallée, Cité Descartes, 5 Boulevard Descartes, Champs-sur-Marne, F-77454 Marne-la-Vallée, France. E-Mail: vandek@univ-mlv.fr

We introduce a new missing-data model, based on a mixture of K Markov processes, and consider the general problem of identifying its parameters. We point out in detail the main difficulties of statistical inference for such models: complete likelihood calculation, parametrization of the stationary distribution and identifiability. We propose a general tractable approach for estimating these models (admitting parametrization of the stationary distribution and identifiability) and check in detail that our assumptions are fully satisfied for a Markov mixture of two linear AR(1) models with Gaussian noise. Finally, a Monte Carlo method is proposed to calculate the split data likelihood of this model when no analytic expression for the invariant probability densities of the Markov processes is known.

Keywords: hidden Markov chain; incomplete data; Markov chain; mixture; statistical inference

1. Introduction

In the signal processing and statistics literatures, different definitions of mixtures of Markov models (MMMs) can be found. In signal processing, the study of MMMs is also associated with the problem of identifying mixed stationary sources, Markovian or not. Let us consider an observable finite sequence of K-dimensional random variables, $X = (X_k)_{1 \leqslant k \leqslant T}$, from an instantaneous mixture of K different sources $S = (S_k)_{1 \leqslant k \leqslant T}$, that is,

$$\forall k = 1, \ldots, T, \qquad X_k = AS_k \tag{1}$$

where A is a square and invertible matrix, called a mixing matrix. The goal in such a framework is to recover the sources S from X by estimating $B = A^{-1}$. See, for example, Pham and Garat (1997) and Dégerine and Zaïdi (2002) for respectively, pseudo- and exact likelihood approaches in the case of mixtures of Gaussian autoregressive (AR) sources, and detailed references.

In statistics – see Jalali and Pemberton (1995), Wong and Li (2000; 2001), Benesch (2001) – the MMMs are defined from a distributional point of view. In the spirit of definition (2.1) for the mixed autoregressive model (MAR) in Wong and Li (2001), we should say that a process $X = (X_n)_{n \geqslant 1}$ is an MMM if, for all $n \geqslant d+1$, $F_{x_n}(x|F_{n-1})$, the

1350–7265

conditional cumulative distribution function of X_n given the past information, takes the form

$$F_{x_n}(x|\mathcal{F}_{n-1}) = \sum_{k=1}^{K} \alpha_k \Phi_k(x; x_{n-1}, \ldots, x_{n-d}; \vartheta_k), \tag{2}$$

with $\sum_{k=1}^{k} \alpha_k = 1$, $\alpha_k > 0$, for $k = 1, \ldots, k$, and where $\Phi_k(\cdot; x_{n-1}, \ldots, x_{n-d}; \vartheta_k)$ is a cumulative distribution function depending on parameters $(x_{n-1}, \ldots, x_{n-d}; \vartheta_k)$ (observations from the past of length d, and a statistical parameter). According to Wong and Li (2000) the MAR is useful for modelling times series with multimodal marginal or conditional distributions; see Tong (1990) and Chan and Tong (1998). An application to real biological data, on the Canadian lynx, is given in the latter two papers. Other models with smooth changes with respect to time have been proposed in econometrics, and later applied to other areas: the so-called autoregressive processes with Markov regime, whose dynamic is driven by a Markov chain. Denoting such a process by X, one basic definition is

$$\forall\, n \geqslant d+1, \qquad X_n = \sum_{i=1}^{d} a_i(U_n) X_{n-i} + \sigma(U_n)\varepsilon_n, \tag{3}$$

where $(\varepsilon)_{n\geqslant 1}$ is sequence of independent and identically distributed (i.i.d.) random variables, $U = (U_n)_{n\geqslant 1}$ is a Markov chain with continuous or discrete state space, and $(a_i(\cdot))_{i=1,\ldots,d}$ and $\sigma(\cdot)$ are functions defined on the state space of U. This model was used by Hamilton (1989) to model US gross national product (the Us modelling the economic/business cycles); see Hamilton and Susmel (1994), Cai (1994) and Garcia and Perron (1996) for recent extensions. Linear autoregressive processes with Markov regime are also widely used in electrical engineering (see Bar-Shalom and Li 1993), failure detection (Tugnait 1982) and automatic control (Ji *et al.* 1990; Krishnamurthy and Rydén 1998). Another important class of autoregressive Markov processes with Markov regime are the hidden Markov models (HMMs), for which the conditional distribution of X_n does not depend on lagged Xs but only on U_n. HMMs are used in many different areas, including speech recognition (see Juang and Rabiner 1991), neurophysiology (Fredkin and Rice 1987) and econometrics (see Chib *et al.* 1998). Most work on maximum likelihood estimation in these models has focused on numerical methods for approximation of the maximum likelihood estimator (MLE). In sharp contrast, it took a long time to achieve significant progress on the statistical issue of the asymptotic properties of the MLE for HMMs and autoregressive processes with Markov regime. On HMMs, see Baum and Petrie (1966), Leroux (1992), Bakry *et al.* (1997), Bickel *et al.* (1998), LeGland and Mevel (2000), Douc and Matias (2001), and on autoregressive processes with Markov regime, see also Krishnamurthy and Rydén (1998) and Francq and Roussignol (1998), when U takes values in a finite set, and Douc *et al.* (2004), when U takes values in a continuous state space.

Let us observe, finally, that probabilistic work on characterization of mixtures of Markov chain distributions was initiated by de Finetti (1959), and has been continued by, among many others, Freedman (1962), Diaconis and Freedman (1980) and more recently Fortini *et al.* (2002).

In this paper we introduce another possible definition of MMMs. Let us consider

$X^{[i]} = (X_n^{[i]})_{n\geqslant 1}$, $1 \leqslant i \leqslant K$, K independent stationary discrete-time Markov processes taking values in a measurable state space $(E, \mathcal{E})$ with probability transition densities Q^i, $1 \leqslant i \leqslant K$, with respect to a common finite dominating measure λ. The MMM we consider induces an observed process $Z = (Z_n)_{n\geqslant 1}$ based on the collection of the K mutually independent processes $(X^{[i]})_{1\leqslant i\leqslant K}$ and defined by,

$$\forall n \geqslant 1, \qquad Z_n = \sum_{i=1}^{K} \mathbf{1}_{\{U_n=i\}} X_n^{[i]}, \tag{4}$$

where $(U_n)_{n\geqslant 0}$ is a stationary positive recurrent Markov chain with values in $\mathcal{U} = \{1, \ldots, K\}$. We suppose, in addition, that the chain $(U_n)_{n\geqslant 0}$ is not observed, which corresponds to a situation where only mixtures of sample paths (due to a Markovian process selection) coming from independent Markov sources are observed. To differentiate this model from other MMMs, we propose to call it the hidden Markov mixture of Markov models (HMMMM or H4M). Let us remark that our MMM is not Markovian and is clearly different from other MMMs. On the other hand, it is worth observing that hidden Markov models belong to the class of H4Ms. To check this point it is enough to consider independent sequences for the $X^{[i]}$ in (4). Notice at this point that HMMs are at the junction of H4Ms and the class of autoregressive models with Markov regime, when the underlying Markov chain U is supposed to belong to a finite state space. From the previous remark the H4Ms are naturally well suited to applications in areas where HMMs hold; recall our observations on HMMs earlier in this section; and see also Section 6.

Having made these preliminary remarks, we wish to draw attention to the ability of our model to describe discrete time series with: (i) abrupt changes, when U undergoes a change of state; (ii) local stationarity, during stages where U remains in the same state; (iii) multimodal marginal distributions from mixture structure; and (iv) *phase-type feedback* effects (see Neuts 1994, p. 46), for the definition of phase-type distributions. Let us elaborate on point (iv). Consider two sample paths of length $n \geqslant 3$: $u_1^n = (u_1, \ldots, u_n)$ from U and $z_1^n = (z_1, \ldots, z_n)$ from Z, and fix $u_{n+1} = i$. Suppose that there exists an index $\tau_n \geqslant 2$ such that $u_{n+1-\tau_n} = i$, $u_{n+1-k} \neq i$ for all $k = 1, \ldots, \tau_n - 1$, that is, corresponding to the time separating the current observation of U at state i and the last observed value of U at state i. From the definition of Z, we can check that the conditional law $\mathcal{L}(Z_{n+1}|U_1^{n+1} = u_1^{n+1}, Z_1^n = z_1^n)$, satisfies

$$\mathcal{L}(Z_{n+1}|U_1^{n+1} = u_1^{n+1}, Z_1^n = z_1^n) = \mathcal{L}(Z_{n+1}|U_{n+1-\tau_n}^{n+1} = u_{n+1-\tau_n}^{n+1}, Z_{n+1-\tau_n} = z_{n+1-\tau_n}), \tag{5}$$

which depends only on i, τ_n and $z_{n-\tau_n}$, from independence of the $X^{[i]}$, and their Markovian structure. Equation (5) shows that the law of the process Z at time $n+1$ is influenced by a particular observation at an unknown time in the past (feedback effect). Let us add a final point: (v) quasi-independence of the homogeneous phases, when U spends long periods in the same states and if the $X^{[i]}$ are strongly mixing.

The goal of this paper is to propose a $\sqrt{n}$-consistent method, based on the maximum split data likelihood estimate (MSDLE) introduced by Rydén (1994), for estimating the parameters driving the transition density kernels of the $X^{[i]}$, and the transition matrix of U. The rest of this paper is organized as follows. In Section 2 we give a precise description of the MSDLE for H4Ms, and the main assumptions. In Section 3 we prove consistency and asymptotic normality of the MSDLE under mild conditions. In Section 4 we propose a Monte Carlo approach to estimate the log of the split data likelihood (SDL), when an analytical expression for the invariant density of the $X^{[i]}$ cannot be given under fixed parametrization of the transitions. Section 5 is devoted to a detailed study of a hidden Markov mixture of two linear autoregressive processes of order 1. In Section 6, we indicate some possible applications for the H4Ms in neurophysiology and kinetics. Also in Section 6 we report sample path simulations of different HMMs and H4Ms with the same marginal distribution and the same underlying Markov chain U. A short empirical comparison of the obtained patterns is made, and similarities with alpha and theta waves found in kinetics are noted.

2. Assumptions and parametrization

For ease of notation, and without loss of generality, we propose to consider the case $K = 2$ and write $X = X^{[1]}$ and $Y = X^{[2]}$. The transition density kernels of X and Y will be parametrized by $\theta \in \Phi^1$ for Q^1 and $\phi \in \Phi^2$ for Q^2, with Φ^i, $i = 1, 2$, compact sets in $\mathbb{R}^q$, and are assumed to belong respectively to the parametric families $\mathcal{K}^1 = \{Q^1_\theta(\cdot, \cdot), \theta \in \Phi^1\}$ and $\mathcal{K}^2 = \{Q^2_\phi(\cdot, \cdot), \phi \in \Phi^2\}$. We suppose that for each $\theta \in \Phi^1$ ($\phi \in \Phi^2$) the probability transition kernel Q^1_θ (Q^2_ϕ) induces a recurrent positive Markov process, and admits a unique invariant probability measure with density q^1_θ (q^2_ϕ). Notice that, in general, analytic expressions of these densities are not explicitly known except in the case of linear autoregressive models with Gaussian noise; see Sections 5 and 6. Nevertheless, let us recall that for each $\theta \in \Phi^1$ and $\phi \in \Phi^2$, q^1_θ and q^2_ϕ are the unique solutions of the functional fixed point problems

$$\int_E q^1_\theta(x_1) Q^1_\theta(x_1, \cdot)\lambda(\mathrm{d}x_1) = q^1_\theta(\cdot), \quad \text{and} \quad \int_E q^2_\phi(y_1) Q^2_\phi(y_1, \cdot)\lambda(\mathrm{d}y_1) = q^2_\phi(\cdot). \tag{6}$$

The transition matrix Π of U will be parametrized by $\gamma = (\alpha, \beta) \in [\delta, 1-\delta]^2$, with $0 < \delta < 1$, as follows:

$$\Pi_\gamma = \begin{pmatrix} \pi_\gamma(1, 1) & \pi_\gamma(1, 2) \\ \pi_\gamma(2, 1) & \pi_\gamma(2, 2) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}. \tag{7}$$

The invariant probability vector associated with Π_γ is denoted by

$$(\pi_\gamma(1), \pi_\gamma(2)) = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right).$$

Finally the *global* parameter which is to be estimated can be written as $\vartheta = (\gamma; \theta, \phi) \in \Theta = [\delta, 1-\delta]^2 \times \Phi^1 \times \Phi^2$. From now on we use the notation $\mathbf{Z}_1^n = \{Z_k;\ 1 \leqslant k \leqslant n\}$ for all processes. Suppose that U is to be observed and consider $u_1^n = (u_1, \ldots, u_n)$ a sample path of length n from U, and $z_1^n = (z_1, \ldots, z_n)$ a sample path of length n from Z. Then the likelihood function for (U, Z) based on (u_1^n, z_1^n) can be written as

$p_\vartheta(u_1^n, z_1^n) = p_\vartheta(z_1^n|u_1^n)p_\vartheta(u_1^n)$, where $p_\vartheta(u_1^n) = P_\vartheta(\mathbf{U}_1^n = u_1^n)$, and $p_\vartheta(z_1^n|u_1^n)$ denotes the density of the Zs conditionally on $\{\mathbf{U}_1^n = u_1^n\}$, which expressions are respectively given by

$$p_\vartheta(u_1^n) = \pi_\gamma(u_1)\prod_{j=1}^{n-1}\pi_\gamma(u_j, u_{j+1})$$

and

$$p_\vartheta(z_1^n|u_1^n) = \int_{E^n} q_\theta^1(x_1)\prod_{j=1}^{n-1}Q_\theta^1(x_j, x_{j+1}) \otimes_{j\in\{1,\ldots,n\}/u_j=2}\lambda(\mathrm{d}x_j)\otimes_{j\in\{1,\ldots,n\}/u_j=1}\delta_{z_j}(\mathrm{d}x_j)$$

$$\times\int_{E^n} q_\phi^2(y_1)\prod_{j=1}^{n-1}Q_\phi^2(y_j, y_{j+1}) \otimes_{j\in\{1,\ldots,n\}/u_j=1}\lambda(\mathrm{d}y_j)\otimes_{j\in\{1,\ldots,n\}/u_j=2}\delta_{z_j}(\mathrm{d}y_j),$$

where we recognize the joint density of the independent random vectors $\mathbf{X}_1^n$ and $\mathbf{Y}_1^n$, integrated componentwise when U is not in state 1 or not in state 2, as appropriate. To compute the likelihood function for the Zs alone, it remains to sum $p_\vartheta(u_1^n, z_1^n)$ over all the possible values of u_1^n, to give

$$p_\vartheta(z_1^n) = \sum_{(u_1,u_2,\ldots,u_n)\in\{1,2\}^n}\pi_\gamma(u_1)\prod_{j=1}^{n-1}\pi_\gamma(u_j, u_{j+1}) \tag{8}$$

$$\times\int_{E^n} q_\theta^1(x_1)\prod_{j=1}^{n-1}Q_\theta^1(x_j, x_{j+1}) \otimes_{j\in\{1,\ldots,n\}/u_j=2}\lambda(\mathrm{d}x_j)\otimes_{j\in\{1,\ldots,n\}/u_j=1}\delta_{z_j}(\mathrm{d}x_j)$$

$$\times\int_{E^n} q_\phi^2(y_1)\prod_{j=1}^{n-1}Q_\phi^2(y_j, y_{j+1}) \otimes_{j\in\{1,\ldots,n\}/u_j=1}\lambda(\mathrm{d}y_j)\otimes_{j\in\{1,\ldots,n\}/u_j=2}\delta_{z_j}(\mathrm{d}y_j).$$

Let us remark that, unlike discrete HMMs, the likelihood of H4Ms does not benefit from a recurrence formula based on the filter, since successive Z_i are not independent conditionally on a finite-length past of U. Surprisingly, this technique allows the otherwise intractable likelihood of HMMs to be computed in linear time (with respect to n); see Rabiner (1989). For this reason, and because of the great complexity of the likelihood function of H4Ms, we propose instead, as a first step, to consider a maximum split data likelihood estimate (MSDLE) in the spirit of Rydén (1994), instead of the highly intractable maximum likelihood estimate (MLE).

For an integer m conveniently chosen, we define the m-dimensional MSDLE based on Z_1^{km}, $k \geqslant r$, as follows

$$\hat{\vartheta}_k = \arg\max_{\vartheta\in\Theta}\prod_{j=1}^{k}p_\vartheta(\mathbf{Z}_{(j-1)m+1}^{jm}). \tag{9}$$

The true parameter value will be denoted by ϑ_0, the law of Z over $E^{\mathbb{N}}$ will be denoted for simplicity by P_0, the index 0 recalling that ϑ_0 entirely defines the law of Z, and expectation under P_0 will be denoted by $\mathrm{E}_0(\cdot)$. The following conditions will be used throughout the paper.

Condition C1. *The true parameter* ϑ_0 *is an interior point of* Θ, *a compact set in* $\mathbb{R}^{2q+2}$.

Condition C2. *The parametric family* $\mathcal{F}^m = \{p_\vartheta(z_1, \ldots, z_m);\ \vartheta \in \Theta\}$ *is identifiable in the sense that*

$$\forall(\vartheta, \vartheta') \in \Theta^2 \mid p_\vartheta(z_1, \ldots, z_m) = p_{\vartheta'}(z_1, \ldots, z_m)\ \lambda^{\otimes m}\text{-almost everywhere} \Rightarrow \vartheta = \vartheta'.$$

Condition C3. *There exist two functions* g_1 *and* g_2 *from* E^m *into* $\mathbb{R}$ *such that*

$$g_1(z_1, \ldots, z_m) \leqslant p_\vartheta(z_1, \ldots, z_m) \leqslant g_2(z_1, \ldots, z_m), \qquad \forall(z_1, \ldots, z_m; \vartheta) \in E^m \times \Theta,$$

and

$$\int_{E^m} |\log(g_i(z_1, \ldots, z_m))| p_{\vartheta_0}(z_1, \ldots, z_m)\lambda(\mathrm{d}z_1^m) < \infty, \qquad i = 1, 2.$$

Condition C4. *The function* $\vartheta \mapsto p_\vartheta(z_1, \ldots, z_m)$ *is* $\lambda^{\otimes m}$*-a.e. twice differentiable on* Θ.

Condition C5. *Write* $\vartheta = (\alpha, \beta; \theta_1, \ldots, \theta_q; \quad \phi_1, \ldots, \phi_q) = (\vartheta_1, \vartheta_2; \quad \vartheta_3, \ldots, \vartheta_{q+2}; \vartheta_{q+3}, \ldots, \vartheta_{2q+2})$, *and let* $\|\cdot\|$ *be the Euclidean norm on* $\mathbb{R}^{2q+2}$. *There exists* $\xi_0 > 0$ *such that:*

(i) *for* $1 \leqslant i \leqslant 2q+2$, *and all* $(z_1, \ldots, z_m) \in E^m$, *there is a function* $g^{(1)}$ *from* E^m *into* $\mathbb{R}$ *such that*

$$\sup_{\|\vartheta-\vartheta_0\|\leqslant\xi_0} \left| \frac{\partial}{\partial\vartheta_i} \log p_\vartheta(z_1, \ldots, z_m) \right| \leqslant g^{(1)}(z_1, \ldots, z_m),$$

with

$$\int_{E^m} g^{(1)}(z_1, \ldots, z_m) p_{\vartheta_0}(z_1, \ldots, z_m)\lambda(\mathrm{d}z_1^m) < \infty,$$

and, for $\kappa > 0$,

$$\int_{E^m} \left(g^{(1)}(z_1, \ldots, z_m)\right)^{2+\kappa} p_{\vartheta_0}(z_1, \ldots, z_m)\lambda(\mathrm{d}z_1^m) < \infty;$$

(ii) *for all* $1 \leqslant i, j \leqslant 2q+2$, *and all* $(z_1, \ldots, z_m) \in E^m$, *there exists a function* $g^{(2)}$ *from* E^m *into* $\mathbb{R}$ *such that*

$$\sup_{\|\vartheta-\vartheta_0\|\leqslant\xi_0} \left| \frac{\partial^2}{\partial^2\vartheta_i\vartheta_j} \log p_\vartheta(z_1, \ldots, z_m) \right| \leqslant g^{(2)}(z_1, \ldots, z_m),$$

and

$$\int_{E^m} g^{(2)}(z_1, \ldots, z_m) p_{\vartheta_0}(z_1, \ldots, z_m)\lambda(\mathrm{d}z_1^m) < \infty.$$

Condition C6. *The partial derivatives of order* 0, 1, 2 *of the function* $\vartheta \mapsto p_\vartheta(z_1, \ldots, z_m)$ *are* $\mathcal{E}^{\otimes m}$*-measurable for each* $\vartheta \in \Theta$.

Condition C7. *The Markov processes X and Y are supposed stationary and geometrically* α*-mixing (or* β*-mixing).*

The definition of α-mixing coefficients for a stationary process is given in (13); see also Doukhan (1994, p. 88) for a simple definition in the case of stationary Markov processes.

3. Consistency and asymptotic normality

In this section we prove under mild conditions that the MSDLE defined in (9) is consistent and asymptotically normal. For this purpose, we begin with a technical lemma useful in treating the asymptotic behaviour of the SDL (and its derivatives).

Lemma 1. (i) *Under Condition C7, for all measurable functions* $\varphi(\cdot)$ *from* E^m *into* $\mathbb{R}^d$, $d \geqslant 1$, *the sequence* $(\varphi(\mathbf{Z}_{(k-1)m+1}^{mk}))_{k\geqslant 1}$ *is stationary and geometrically* α*-mixing.*

(ii) *Under the assumptions of* (i), *for all* $\varphi \in L_1(P_0)$ *we have the strong law of large numbers, that is,*

$$M_k = \frac{1}{k}\sum_{j=1}^{k} \varphi(\mathbf{Z}_{(j-1)m+1}^{jm}) \underset{k\to\infty}{\longrightarrow} E_0(\varphi(\mathbf{Z}_1^m)), \quad P_0\text{-almost surely.} \tag{10}$$

(iii) *Suppose that* $\mathrm{E}(\varphi(\mathbf{Z}_1^m)) = 0$, $\mathrm{E}|\varphi(\mathbf{Z}_1^m)|^{2+\kappa} < \infty$ *for some* $\kappa > 0$, *and that Condition C7 is satisfied. Then*

$$\Sigma \overset{def.}{=} \mathrm{E}(\varphi(\mathbf{Z}_1^m)^2) + 2\sum_{k=1}^{\infty} k\mathrm{E}(\varphi(\mathbf{Z}_1^m)\varphi^{\mathrm{T}}(\mathbf{Z}_{(k-1)m+1}^{mk})) < \infty, \tag{11}$$

and, if $\Sigma \neq 0_{d\times d}$,

$$\sqrt{k}M_k \underset{k\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0, \Sigma). \tag{12}$$

Proof. (i) Without loss of generality, we consider the case $m = 2$. The function $\varphi(\cdot)$ being $\mathcal{E}^{\otimes 2}$-measurable, it is enough to consider the α-mixing coefficient associated with the Markov process $W = (W_k)_{k\geqslant 1} = (\mathbf{X}_{(k-1)2+1}^{2k}, \mathbf{Y}_{(k-1)2+1}^{2k}, \mathbf{U}_{(k-1)2+1}^{2k})_{k\geqslant 1}$. At this stage, let us define for all stationary processes $\tilde{X}$, and all $(t, k) \in \mathbb{N}^* \times \mathbb{N}$, the sequence of α-mixing coefficients associated with $\tilde{X}$ by

$$\alpha^{\tilde{X}}(k) = \sup_{A\in\mathcal{F}_{\tilde{X},1}^{t}, B\in\mathcal{F}_{\tilde{X},t+k+1}^{\infty}} |P(A\cap B) - P(A)P(B)|, \tag{13}$$

where $\mathcal{F}_{\tilde{X},t_1}^{t_2}$ denotes, for all $t_2 > t_1 \geqslant 1$, the σ-algebra generated by $(\tilde{X}_{t_1}, \ldots, \tilde{X}_{t_2})$. Following (13), the sequence of α-mixing coefficients associated with W is defined by

$$\alpha^W(k) = \sup |P((A_1, A_2, A_U) \cap (B_1, B_2, B_U)) - P(A_1, A_2, A_U)P(B_1, B_2, B_U)|, \tag{14}$$

where the supremum is taken over all $(A_1, A_2, A_U) \in \mathcal{F}_{X,1}^{2n} \otimes \mathcal{F}_{Y,1}^{2n} \otimes \mathcal{F}_{U,1}^{2n}$ and $(B_1, B_2, B_U) \in \mathcal{F}_{X,2n+2k+1}^{\infty} \otimes \mathcal{F}_{Y,2n+2k+1}^{\infty} \otimes \mathcal{F}_{U,2n+2k+1}^{\infty}$.

From the mutual independence of X, Y and U, the modulus of the difference of probabilities in the right-hand side of (14) satisfies

$$\begin{aligned}
&|P(A_1 \cap B_1)P(A_2 \cap B_2)P(A_U \cap B_U) - P(A_1)P(B_1)P(A_2)P(B_2)P(A_U)P(B_U)| \\
&\quad = P(A_U)P(A_1)P(A_2)|P(B_1|A_1)P(B_2|A_2)[P(B_U|A_U) - P(B_U)] \\
&\qquad + P(B_1|A_1)P(B_U)[P(B_2|A_2) - P(B_2)] \\
&\qquad + P(B_2)P(B_U)[P(B_1|A_1) - P(B_1)]| \\
&\quad \leqslant |P(A_U \cap B_U) - P(A_U)P(B_U)| + |P(A_1 \cap B_1) - P(A_1)P(B_1)| \\
&\qquad + |P(A_2 \cap B_2) - P(A_2)P(B_2)|.
\end{aligned}$$

From the Markovian structure of U, X and Y, Condition C7, and the last inequality, we obtain

$$\alpha^W(k) \leqslant \alpha^U(2k) + \alpha^X(2k) + \alpha^Y(2k) \leqslant \rho^k, \tag{15}$$

for a certain $0 < \rho < 1$, and k large enough. Notice now that the mapping s from $E^2 \times E^2 \times \{1, 2\}^2$ into E^2, such that $\mathbf{Z}_1^2 = s(\mathbf{X}_1^2; \mathbf{Y}_1^2; \mathbf{U}_1^2)$ – see (4) – and defined by

$$s(x_1, x_2; y_1, y_2; u_1, u_2) = ((2 - u_1)x_1 + (u_1 - 1)y_1; (2 - u_2)x_2 + (u_2 - 1)y_2),$$

is measurable, hence $\varphi \circ s$ is a measurable function from $E^2 \times E^2 \times \{1, 2\}^2$ into $\mathbb{R}^d$, which means that the α-mixing coefficients of the sequence $(\varphi(\mathbf{Z}_{(k-1)m+1}^{mk}))_{k\geqslant 1}$ are inferior or equal to the coefficients induced by W, which, using (15), concludes the proof of (i).

(ii) This result is a direct consequence of the maximal ergodic theorem for stationary processes; see Stout (1974, p. 145).

(iii) This central limit theorem is a classical result, see (29.10) in Billingsley (1995, p. 387), which is proved by considering the central limit theorem for real α-mixing sequences of random variables, see Theorem 3.2.1 in Zhengyan and Chuanrong (1996), and the Cramér–Wold device, see Theorem 29.4 in Billingsley (1995, p. 383). □

Theorem 1. *Under Conditions C1–C7, the MSDLE defined in* (9) *is strongly consistent, that is,*

$$\hat{\vartheta}_k \underset{k\to\infty}{\longrightarrow} \vartheta_0 \quad P_0\text{-}a.s., \tag{16}$$

where ϑ_0 is the true value of the parameter.

Proof. The proof is based on the proof given by Dacunha-Castelle and Duflo (1993, pp. 94–96). First of all, the MSDLE can be defined as a minimum contrast estimator,

$$\hat{\vartheta}_k = \arg\min_{\vartheta \in \Theta} U_k(\vartheta),$$

where

$$U_k(\vartheta) = -k^{-1}\ell_\vartheta(\mathbf{Z}_1^{mk}), \qquad \ell_\vartheta(\mathbf{Z}_1^{mk}) = \log \prod_{j=1}^{k} p_\vartheta(\mathbf{Z}_{(j-1)m+1}^{mj}) = \sum_{j=1}^{k} \log p_\vartheta(\mathbf{Z}_{(j-1)m+1}^{mj}), \quad (17)$$

where $\ell_\vartheta(z_1^{km})$ denotes the log of the SDL. From Lemma 1, Conditions C3 and C7, we obtain

$$U_k(\vartheta) = -k^{-1}\ell_\vartheta(\mathbf{Z}_1^{mk}) \underset{k\to\infty}{\longrightarrow} \mathcal{E}_0(\vartheta) = -\mathrm{E}_0(\log p_\vartheta(\mathbf{Z}_1^{m})), \quad P_0\text{-a.s.} \quad (18)$$

with $|\varepsilon_0(\vartheta)| < \infty$, for all $\vartheta \in \Theta$.

Under Condition C2 it is clear from Jensen inequality and Condition C2, that

$$\mathcal{E}_0(\vartheta_0) \leqslant \mathcal{E}_0(\vartheta) \quad \text{and} \quad \mathcal{E}_0(\vartheta) = \mathcal{E}_0(\vartheta_0) \Rightarrow \vartheta = \vartheta_0.$$

We now consider the Kullback distance $K(\vartheta_0, \vartheta) = \mathcal{E}_0(\vartheta_0) - \mathcal{E}_0(\vartheta) \geqslant 0$, with $K(\vartheta_0, \vartheta) = 0 \Leftrightarrow \vartheta_0 = \vartheta$. Let us consider D a countable dense set in Θ, so that $\inf_{\vartheta\in\Theta} U_k(\vartheta) = \inf_{\vartheta\in\Theta\cap D} U_k(\vartheta)$ is an $\mathcal{F}_{Z,1}^{k}$-measurable random variable. We define, in addition, the random variable

$$W(k, \eta) = \sup\{|U_k(\vartheta) - U_k(\vartheta')|; (\vartheta, \vartheta') \in D^2, \ |\vartheta - \vartheta'| \leqslant \eta\},$$

and recall that $\mathcal{K}(\vartheta_0, \vartheta_0) = 0$. Let us consider a non-empty open ball B_0 centred in ϑ_0 such that $K(\vartheta_0, \vartheta)$ is bounded from below by a positive real number 2ε on $\Theta\backslash B_0$. Let us consider a sequence $(\eta_r)_{r\geqslant 0}$ decreasing towards zero, and cover $\Theta\backslash B_0$ by a finite number ℓ of balls $(B_i)_{1\leqslant i\leqslant \ell}$, respectively centred in $(\vartheta_i)_{1\leqslant i\leqslant \ell}$, and of radius less than η_r for one r fixed arbitrarly. For all $\vartheta \in B_i$, then,

$$\begin{aligned} U_k(\vartheta) &\geqslant U_k(\vartheta_i) - |U_k(\vartheta) - U_k(\vartheta_i)| \\ &\geqslant U_k(\vartheta_i) - \sup_{\vartheta\in B_i}|U_k(\vartheta) - U_k(\vartheta_i)|, \end{aligned}$$

which leads to

$$\inf_{\vartheta\in\Theta\backslash B_0} U_k(\vartheta) \geqslant \inf_{1\leqslant i\leqslant \ell} U_k(\vartheta_i) - W(k, \eta_r).$$

As a consequence, we have the following event inclusions:

$$
\begin{aligned}
\left\{\hat{\vartheta}_k \notin B_0\right\} &\subseteq \left\{\inf_{\vartheta\in\Theta\setminus B_0} U_k(\vartheta) < \inf_{\vartheta\in B_0} U_k(\vartheta)\right\}\\
&\subseteq \left\{\inf_{\vartheta\in\Theta\setminus B_0} U_k(\vartheta) < U_k(\vartheta_0)\right\}\\
&\subseteq \left\{\inf_{1\leqslant i\leqslant \ell} U_k(\vartheta_i) - W(k, \eta_r) < U_k(\vartheta_0)\right\}\\
&\subseteq \{W(k, \eta_r) > \varepsilon\} \cup \left\{\inf_{1\leqslant i\leqslant \ell} (U_k(\vartheta_i) - U_k(\vartheta_0)) \leqslant \varepsilon\right\}.
\end{aligned}
$$

Thus we have

$$
\limsup_k \left\{\hat{\vartheta}_k \notin B_0\right\} \subseteq \limsup_k \{W(k, \eta_r) > \varepsilon\} \cup \limsup_k \left\{\inf_{1\leqslant i\leqslant \ell} (U_k(\vartheta_i) - U_k(\vartheta_0)) \leqslant \varepsilon\right\}. \tag{19}
$$

By the strong law of large number established in (18) we have

$$
P_0\left(\limsup_k \left\{\inf_{1\leqslant i\leqslant \ell} (U_k(\vartheta_i) - U_k(\vartheta_0)) \leqslant \varepsilon\right\}\right) = 0. \tag{20}
$$

In addition, according to Condition C3, there exists a random variable $h(\mathbf{Z}_1^m)$ such that

$$
\sup_{\vartheta\in\Theta} |\log p_\vartheta(\mathbf{Z}_1^m)| \leqslant h(\mathbf{Z}_1^m),
$$

with $\mathrm{E}_0[h(\mathbf{Z}_1^m)] < \infty$, where $h = |\log g_1| + |\log g_2|$ does not depend on ϑ. Let us consider the random variable

$$
H_\eta\left(\mathbf{Z}_1^{\mathrm{m}}\right) = \sup_{(\vartheta,\vartheta')\in\Theta^2} \left\{|\log p_\vartheta\left(\mathbf{Z}_1^m\right) - \log p_{\vartheta'}\left(\mathbf{Z}_1^m\right)|;\ |\vartheta - \vartheta'| \leqslant \eta\right\}.
$$

Using the previous uniform upper bound and continuity Condition C4, we obtain that

$$
H_\eta\left(\mathbf{Z}_1^m\right) \leqslant 2h\left(\mathbf{Z}_1^m\right) \quad \text{and} \quad \lim_{\eta\to 0} E_0\left[H_\eta\left(\mathbf{Z}_1^m\right)\right] = 0.
$$

Hence, for r' large enough, we have $\mathrm{E}_0(H_{\eta_{r'}}(\mathbf{Z}_1^m)) \leqslant \varepsilon$, and $W(k, \eta_{r'}) \leqslant k^{-1}\sum_{j=1}^k H_{\eta_{r'}}(\mathbf{Z}_{(j-1)m+1}^{mj})$ P_0-almost surely; therefore,

$$
\limsup_k \{W(k, \eta_{r'}) > \varepsilon\} \subseteq \limsup_k \left\{k^{-1}\sum_{j=1}^k H_{\eta_{r'}}(\mathbf{Z}_{(j-1)m+1}^{mj}) > \varepsilon\right\},
$$

and

$$
P_0\left(\limsup_k \left\{k^{-1}\sum_{j=1}^k H_{\eta_{r'}}(\mathbf{Z}_{(j-1)m+1}^{mj}) > \varepsilon\right\}\right) = 0
$$

which leads to

$$P_0\left(\limsup_k\{W(k,\eta_{r'})>\varepsilon\}\right)=0. \tag{21}$$

By (19)–(21), we prove the strong consistency of the MSDLE $\hat{\vartheta}_k$. □

Write $V_j(\vartheta)=\log p_\vartheta(\mathbf{Z}_{(j-1)m+1}^{mj})$ for $j=1,\ldots,k$, and let us denote for any function v depending on ϑ, its gradient vector and Hessian matrix respectively by

$$\dot{v}(\vartheta)=\frac{\partial v}{\partial\vartheta}(\vartheta)\quad\text{and}\quad\ddot{v}(\vartheta)=\frac{\partial^2 v}{\partial\vartheta\partial\vartheta^{\mathrm{T}}}(\vartheta). \tag{22}$$

Let us write $\ell_k(\vartheta)=\ell_\vartheta(\mathbf{Z}_1^{2k})$, in line with the notation in (22). From (17) we obtain

$$k^{-1/2}\dot{\ell}_k(\vartheta)=k^{-1/2}\sum_{j=1}^{k}\dot{V}_j(\vartheta). \tag{23}$$

Lemma 2. *Under Conditions C4–C7, we have*

$$k^{-1/2}\dot{\ell}_k(\vartheta_0)\ \xrightarrow[k\to\infty]{\mathcal{L}}\ \mathcal{N}(0,\Sigma_0),$$

where

$$\Sigma_0=\mathrm{E}_0(\dot{V}_1(\vartheta_0)\dot{V}_1^{\mathrm{T}}(\vartheta_0))+2\sum_{k=2}^{\infty}k\,\mathrm{E}_0(\dot{V}_1(\vartheta_0)\dot{V}_k^{\mathrm{T}}(\vartheta_0))<\infty.$$

Proof. This result is a direct consequence of Conditions C4–C7 and Lemma 1, taking $\varphi(\cdot)=\partial\log p_{\vartheta_0}(\cdot)/\partial\vartheta$. □

Lemma 3. *Let $(\vartheta_k^*)_{k\geqslant 0}$ be any arbitrary sequence converging P_0-a.s. towards ϑ_0. Under Conditions C4–C7, we have*

$$k^{-1}\ddot{\ell}_k(\vartheta_k^*)\ \xrightarrow[k\to\infty]{}\ A_0=\mathrm{E}_0(\ddot{V}_1(\vartheta_0))$$

in P_0-probability.

Proof. By Lemma 1 and Condition C4 we know that $(k^{-1}\ddot{\ell}_k(\vartheta_0))_{k\geqslant 1}$ converges P_0-a.s. to $\mathrm{E}_0\left[\ddot{V}_1(\vartheta_0)\right]$. Now let us prove that $(k^{-1}\ddot{\ell}_k(\vartheta_k^*))_{k\geqslant 1}$ and $(k^{-1}\ddot{\ell}_k(\vartheta_0))_{k\geqslant 1}$ are asymptotically equivalent in P_0-probability, that is,

$$\forall\eta>0,\qquad\lim_{k\to\infty}P_0\left(\left|\frac{1}{k}\ddot{\ell}_k(\vartheta_k^*)-\frac{1}{k}\ddot{\ell}_k(\vartheta_0)\right|>\eta\right)=0,$$

where we denote (in this proof) by $|\cdot|$ the norm on the real matrices defined, for all $d\times d$ real matrices $A=(A_{i,j})_{i,j=1,\ldots,d}$, by $|A|=\max_{i,j=1,\ldots,d}|A_{i,j}|$, with the convention that the norm of a scalar coincides with its modulus. For this purpose we notice that for all $0<\xi<\xi_0$ (for definition of ξ_0, see Condition C5), we can write

$$P_0\left(\left|\frac{1}{k}\ddot{\ell}_k(\vartheta_k^*) - \frac{1}{k}\ddot{\ell}_k(\vartheta_0)\right| > \eta\right) \leqslant P_0\left(\frac{1}{k}\sum_{j=1}^{k}\sup_{\vartheta\in B(\vartheta_0,\xi)}\left|\ddot{V}_j(\vartheta) - \ddot{V}_j(\vartheta_0)\right| > \eta\right)$$
$$+ P_0(\vartheta_k^* \notin B(\vartheta_0, \xi)),$$

where $B(\vartheta_0, \xi)$ denotes the ball centred on ϑ_0, with radius equal to ξ. The second term on the right-hand side goes to zero as k goes to infinity by strong consistency of ϑ_k^*. For the first term on the right-hand side we notice that

$$\varrho(\xi; \mathbf{Z}_1^m) = \sup_{\vartheta\in B(\vartheta_0,\xi)}\left|\ddot{V}_j(\vartheta) - \ddot{V}_j(\vartheta_0)\right| \underset{\xi\to 0}{\longrightarrow} 0 \text{ a.e.}$$

In addition there exists, from Condition C5, a P_0-integrable function $g^{(2)}$, such that, for all $j = 1, \ldots, k$, the components of the matrix $\ddot{V}_j(\vartheta)$ are all dominated in modulus by $g^{(2)}(\mathbf{Z}_1^m)$ on $B(\vartheta_0, \xi_0)$, which implies that $\varrho(\xi; \mathbf{Z}_1^m) \leqslant 2g^{(2)}(\mathbf{Z}_1^m)$. Now, using the Lebesgue continuity theorem, we obtain that

$$\mathrm{E}_0\left[\varrho(\xi; \mathbf{Z}_1^m)\right] \underset{\xi\to 0}{\longrightarrow} 0. \tag{24}$$

For all $\varepsilon > 0$, and all $\xi > 0$ small enough such that $0 < \mathrm{E}_0\left[\varrho(\xi; \mathbf{Z}_1^m)\right] < \varepsilon$, we have, using Chebyshev's inequality for positive random variables,

$$P_0\left(\frac{1}{k}\sum_{j=1}^{k}\varrho(\xi; \mathbf{Z}_{(j-1)m+1}^{mj}) \geqslant \varepsilon\right) \leqslant P_0\left(\frac{1}{k}\sum_{j=1}^{k}\varrho(\xi; \mathbf{Z}_{(j-1)m+1}^{mj}) \geqslant \varepsilon - \mathrm{E}_0\left[\varrho(\xi; \mathbf{Z}_1^m)\right]\right)$$
$$\leqslant \frac{1}{k[\varepsilon - \mathrm{E}_0[\varrho(\xi; \mathbf{Z}_1^m)]]}\sum_{j=1}^{k}\mathrm{E}_0\left[\varrho(\xi; \mathbf{Z}_{(j-1)m+1}^{mj})\right]$$
$$= \frac{1}{\varepsilon - \mathrm{E}_0\left[\varrho(\xi; \mathbf{Z}_1^m)\right]}\mathrm{E}_0\left[\varrho(\xi; \mathbf{Z}_1^m)\right],$$

which goes to zero, by (24), as ξ goes to 0. □

Theorem 2. *Under Conditions C1–C7, and assuming that A_0 is non-singular, we obtain that*

$$k^{1/2}(\hat{\vartheta}_k - \vartheta_0) \underset{k\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0, A_0^{-1}\Sigma_0 A_0^{-1}).$$

Proof. For k large enough $\hat{\vartheta}_k$ is an interior point of Θ, and $\|\hat{\vartheta}_k - \vartheta_0\| < \xi_0$, and then by a Taylor expansion of $\dot{\ell}_\vartheta(\mathbf{Z}_1^{km})$ about ϑ_0 we obtain,

$$k^{1/2}(\hat{\vartheta}_k - \vartheta_0) = [-k^{-1}\ddot{\ell}_k(\vartheta_k^*)]^{-1}k^{-1/2}\dot{\ell}_k(\vartheta_0),$$

where ϑ_k^* is a point on the line segment between ϑ_0 and $\hat{\vartheta}_k$. Therefore, using Theorem 1 and Lemmas 2 and 3, we obtain the asymptotic normality of the MSDLE. □

4. Monte Carlo estimate of the log of the SDL

We noticed in Section 2 that, except in the case of linear Gaussian autoregressive models, the invariant probability densities q_θ and q_ϕ involved in (8) are analytically unknown, but are solutions of the fixed point problems in (6). The goal of this section is to propose a general tractable approach to approximating the log of the SDL (see (17)), for a given fixed sample path z_1^{mk}, $k \geqslant 1$. The methodology presented here is inspired by Chauveau and Vandekerkhove (2001). Let consider for simplicity the case $k = 1$. In this framework the crucial point is to estimate numerically, for each $\theta \in \Phi^1$ and each $\phi \in \Phi^2$, the quantities $q_\theta^1(z_1)$ and $q_\phi^2(z_1)$. We illustrate our method only on $q_\theta^1(z_1)$, the same procedure holding for $q_\phi^2(z_1)$ (the associated estimate will be denoted by $\hat{q}_\phi^2(z_1)$. Let suppose that, for each $\theta \in \Phi^1$, we are able to simulate an ergodic Markov process $X^\theta = (X_N^\theta)_{N\geqslant 1}$, from Q_θ^1 (knowledge of the stationary initial distribution is not needed in practice, a long burn-in of the chain suffices). From the strong law of large numbers for ergodic Markov chains, and from (6), we obtain

$$\hat{q}_\theta^1(z_1) = \frac{1}{N}\sum_{i=1}^{N} Q_\theta^1(X_i^\theta, z_1) \underset{N\to\infty}{\longrightarrow} \int_E q_\theta^1(x) Q_\theta^1(x, z_1)\lambda(\mathrm{d}x) = q_\theta^1(z_1), \qquad P_\theta\text{-a.s.}$$

Hence $\hat{q}_\theta^1(z_1)$ is a strongly convergent estimate of $q_\theta^1(z_1)$. From this, it is easy to construct a consistent plug-in estimator $\hat{\ell}_\vartheta(z_1^m)$ of $\ell_\vartheta(z_1^m)$, replacing $q_\theta^1(z_1)$ and $q_\phi^2(z_1)$ respectively by $\hat{q}_\theta^1(z_1)$ and $\hat{q}_\phi^2(z_1)$ in (8). In addition, the central limit theorem for $\hat{\ell}_\vartheta(z_1^m)$ can be established. Write, for simplicity,

$$\hat{\ell}_\vartheta(z_1^m) = \log(\hat{q}_\theta^1(z_1)c_1 + \hat{q}_\phi^2(z_1)c_2) \quad \text{and} \quad \ell_\vartheta(z_1^m) = \log(q_\theta^1(z_1)c_1 + q_\phi^2(z_1)c_2),$$

where c_1 and c_2 are constants depending on z_2^m and ϑ. Let us suppose a Condition C7$'$ equivalent to Condition C7, but true for all $\vartheta \in \Theta$ (and not only for ϑ_0). Then, supposing moments conditions akin to Lemma 1 on $Q_\theta^1(z_1, \cdot)$ and $Q_\phi^2(z_1, \cdot)$,

$$\sqrt{N}(\hat{q}_\theta^1(z_1) - q_\theta^1(z_1)) \underset{N\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, \Sigma^1), \quad \text{and} \quad \sqrt{N}(\hat{q}_\phi^2(z_1) - q_\phi^2(z_1)) \underset{N\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, \Sigma^2). \tag{25}$$

where Σ^1 and Σ^2 are variance terms similar to Σ defined in (11). Finally by a Taylor expansion of the log function about $q_\theta^1(z_1)c_1 - q_\phi^2(z_1)c_2$, we have, for all $N \geqslant 1$:

$$\sqrt{N}(\hat{\ell}_\vartheta(z_1^m) - \ell_\vartheta(z_1^m)) = \frac{1}{\ell_N^*}(\sqrt{N}(\hat{q}_\theta^1(z_1) - q_\theta^1(z_1))c_1 + \sqrt{N}(\hat{q}_\phi^2(z_1) - q_\phi^2(z_1))c_2),$$

where ℓ_N^* is a point on the line between $\hat{q}_\theta^1(z_1)c_1 - \hat{q}_\phi^2(z_1)c_2$ and $q_\theta^1(z_1)c_1 - q_\phi^2(z_1)c_2$. From (25) and convergence in P_ϑ-probability of ℓ_n^* towards $q_\theta^1(z_1)c_1 - q_\phi^2(z_1)c_2$, we obtain

$$\sqrt{N}(\hat{\ell}_\vartheta(z_1^m) - \ell_\vartheta(z_1^m)) \underset{N\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, \Sigma),$$

where $\Sigma = [q_\theta^1(z_1)c_1 - q_\phi^2(z_1)c_2]^{-2}[c_1^2\Sigma^1 + c_2^2\Sigma^2]$. In conclusion, we have proposed a $\sqrt{N}$-consistent method to calculate the terms of the form $\log p_\vartheta(z_{(j-1)m+1}^{mj})$, $1 \leqslant j \leqslant k$, and hence for k fixed, and for each $\vartheta \in \Theta$ a $\sqrt{N}$-consistent method exists to calculate the log of the SDL (see (17)) at z_1^{mk} (the asymptotic variance growing linearly with k). From a practical point of view, this approach at least enables the log of the SDL to be computed pointwise, and a discretized version of the MSDLE to be implemented on a grid over Θ.

5. Hidden Markov mixture of two AR(1)s

In this section we check in detail that our mixing, identifiability, regularity and integrability Conditions (C2–C7) in Section 2 are satified for a hidden Markov mixture of two autoregressive processes of order 1. More precisely, the processes X and Y considered in this section are defined, for all $n \geqslant 1$, by

$$X_{n+1} = a_1 X_n + \varepsilon_{n+1} \quad \text{and} \quad Y_{n+1} = a_2 Y_n + \varepsilon'_{n+1}, \tag{26}$$

where $(a_1, a_2) \in (0, 1)$, $(\varepsilon_n)_{n\geqslant 1}$ and $(\varepsilon'_n)_{n\geqslant 1}$ are two mutually independent sequences of independent Gaussian random variables with respective means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. The mixture process U is a Markov chain on $\{1, 2\}$ with transition matrix defined in (7). The parameter in such a set up is

$$\vartheta = (\alpha, \beta, \mu_1, \mu_2, a_1, a_2, \sigma_1^2, \sigma_2^2).$$

We do not describe precisely at this stage the form of the parametrical space Θ since it will be deduced from the coming discussion about identifiability.

Mixing. Condition C7 is clearly satisfied for U, and the same holds for X and Y since these processes are geometrically β-mixing (and hence α-mixing); see, for example, Baraud *et al.* (2001) for general conditions.

Identifiability. Processes X and Y, defined in (26), with initial conditions x_1 and y_1, satisfy

$$X_{n+1} = a_1^n x_1 + \sum_{k=0}^{n-1} a_1^k \varepsilon_{n+1-k}, \qquad Y_{n+1} = a_2^n y_1 + \sum_{k=0}^{n-1} a_2^k \varepsilon'_{n+1-k}. \tag{27}$$

From these expressions and properties of Gaussian random vectors, it is easy to identify the density of the stationary distribution of X and Y processes. In fact for X we obtain the density of a $\mathcal{N}(m_1, s_1)$ distribution, while for Y we have the density of a $\mathcal{N}(m_2, s_2)$ distribution, with

$$m_1 = \frac{\mu_1}{1-a_1}, \quad s_1 = \frac{\sigma_1^2}{1-a_1^2}, \qquad m_2 = \frac{\mu_2}{1-a_2}, \quad s_2 = \frac{\sigma_2^2}{1-a_2^2}. \tag{28}$$

In order to prove identifiability, we propose to consider $m = 2$, and to check that for all ϑ and ϑ' in Θ (which needs to be defined), we have:

$$p_\vartheta(z_1, z_2) = p_{\vartheta'}(z_1, z_2),\ \lambda^{\otimes 2}\text{-a.e.} \Rightarrow \vartheta = \vartheta'. \tag{29}$$

Nethertheless partial information on identifiability will be given considering the marginal equality

$$p_\vartheta(z_2) = \int_{\mathbb{R}} p_\vartheta(z_1, z_2)\mathrm{d}z_1 = \int_{\mathbb{R}} p_{\vartheta'}(z_1, z_2)\mathrm{d}z_1 = p_{\vartheta'}(z_2),\ \lambda\text{-a.e.} \tag{30}$$

Denoting by f_{μ,σ^2} the density function of a $\mathcal{N}(\mu, \sigma^2)$ distribution, we have, for all $\vartheta \in \Theta$:

$$p_\vartheta(z_2) = \pi_\vartheta(1) f_{(m_1,s_1)}(z_2) + \pi_\vartheta(2) f_{(m_2,s_2)}(z_2). \tag{31}$$

Teicher (1963) establishes identifiability property for mixtures of various density families. A mixture of at most r elements of $\mathcal{G} = \{g(z;\,\theta);\ \theta \in \Phi\}$ is identifiable if, for θ_i and θ'_i, for $i = 1, \ldots, r$, in Φ, $(c_1, \ldots, c_r)$ and $(c'_1, \ldots, c'_r)$ probability vectors we have

$$\sum_{i=1}^{r} c_i g(z;\,\theta_i) = \sum_{i=1}^{r} c'_i g(z;\,\theta'_i),\ \lambda\text{-a.e.} \Rightarrow \sum_{i=1}^{r} c_i \delta_{\theta_i} = \sum_{i=1}^{r} c'_i \delta_{\theta'_i}.$$

(δ_θ denotes the point mass at θ). This argument is equivalent to the following statement: there exists a unique permutation σ on $\{1, \ldots, r\}$ such that for all $i = 1, \ldots, r$, $(c_i, \theta_i) = (c'_{\sigma(i)}, \theta'_{\sigma(i)})$. Teicher (1963) shows in particular that mixtures of Gaussian densities (where $\theta_i = (m_i, s_i)$, for $i = 1, \ldots, r$, with m_i denoting the mean parameter, and s_i denoting the variance parameter of the ith component of the mixture) are identifiable, and propose to order the parameter space to avoid the previous permutation ambiguities, that is, by imposing $m_1 < \ldots < m_r$ if $s_i = s_j$, for all $i, j = 1, \ldots, r$, or $(m_i, s_i) < (m_j, s_j)$ if $s_i < s_j$ or $m_i < m_j$ if $s_i = s_j$. In practice, equality of the variance parameters is assumed in order to obtain a simple parameter space for $(m_1, \ldots, m_r)$. The same approach can be used on $(s_1, \ldots, s_r)$ if the variance parameters are assumed all different (without any order constraints on $(m_1, \ldots, m_r)$). Imposing that $s_1 < s_2$ in Θ, the mixture equality (30) and writing (31), we obtain a first partial identification

$$\pi_\vartheta(1) = \pi_{\vartheta'}(1), \quad \pi_\vartheta(2) = \pi_{\vartheta'}(2), \qquad m_1 = m'_1, \quad m_2 = m'_2, \qquad s_1 = s'_1, \quad s_2 = s'_2. \tag{32}$$

For simplicity let us denote $\pi(\cdot) = \pi_\vartheta(\cdot)$, $\pi'(\cdot) = \pi_{\vartheta'}(\cdot)$, $\pi(\cdot,\cdot) = \pi_\vartheta(\cdot,\cdot)$ and $\pi'(\cdot,\cdot) = \pi_{\vartheta'}(\cdot,\cdot)$. Using this first identification in (29), we obtain the following relation which is to be discussed:

$$\begin{aligned}
&\pi(1)\pi(1,1) f_{(m_1,s_1)}(z_1) f_{(a_1z_1+\mu_1,\sigma_1^2)}(z_2) + \pi(1)\pi(1,2) f_{(m_1,s_1)}(z_1) f_{(m_2,s_2)}(z_2)\\
&\quad + \pi(2)\pi(2,1) f_{(m_2,s_2)}(z_1) f_{(m_1,s_1)}(z_2) + \pi(2)\pi(2,2) f_{(m_2,s_2)}(z_1) f_{(a_2z_1+\mu_2,\sigma_2^2)}(z_2)\\
&= \pi(1)\pi'(1,1) f_{(m_1,s_1)}(z_1) f_{(a'_1z_1+\mu'_1,\sigma'^2_1)}(z_2) + \pi(1)\pi'(1,2) f_{(m_1,s_1)}(z_1) f_{(m_2,s_2)}(z_2)\\
&\quad + \pi(2)\pi'(2,1) f_{(m_2,s_2)}(z_1) f_{(m_1,s_1)}(z_2) + \pi(2)\pi'(2,2) f_{(m_2,s_2)}(z_1) f_{(a'_2z_1+\mu'_2,\sigma'^2_2)}(z_2).
\end{aligned} \tag{33}$$

Taking the Fourrier transform term by term with respect to z_2, we obtain

$$\begin{aligned}
&\pi(1)\pi(1,1) f_{(m_1,s_1)}(z_1) \mathrm{e}^{\mathrm{i}t(a_1z_1+\mu_1)-\sigma_1^2t^2} + \pi(1)\pi(1,2) f_{(m_1,s_1)}(z_1) \mathrm{e}^{\mathrm{i}tm_2-s_2t^2}\\
&\quad + \pi(2)\pi(2,1) f_{(m_2,s_2)}(z_1) \mathrm{e}^{\mathrm{i}tm_1-s_1t^2} + \pi(2)\pi(2,2) f_{(m_2,s_2)}(z_1) \mathrm{e}^{\mathrm{i}t(a_2z_1+\mu_2)-\sigma_2^2t^2}\\
&= \pi(1)\pi'(1,1) f_{(m_1,s_1)}(z_1) \mathrm{e}^{\mathrm{i}t(a'_1z_1+\mu'_1)-\sigma'^2_1t^2} + \pi(1)\pi'(1,2) f_{(m_1,s_1)}(z_1) \mathrm{e}^{\mathrm{i}tm_2-s_2t^2}\\
&\quad + \pi(2)\pi'(2,1) f_{(m_2,s_2)}(z_1) \mathrm{e}^{\mathrm{i}tm_1-s_1t^2} + \pi(2)\pi'(2,2) f_{(m_2,s_2)}(z_1) \mathrm{e}^{\mathrm{i}t(a'_2z_1+\mu'_2)-\sigma'^2_2t^2}.
\end{aligned} \tag{34}$$

Let us consider the case $\sigma_1^2 \neq \sigma_2^2$. We begin with the subcase $\sigma_2^2 < \sigma'^2_2 < s_2$, $\sigma_1^2 < \sigma'^2_1 < s_1$

and $\sigma_2^2 < \sigma_1^2$. Multiplying both sides of (34) by $e^{\sigma_2 t^2}$, and taking the limit as t goes to infinity, we obtain the absurd result $\pi(2, 2) = 0$. Let us consider the more complicated subcase $s_1 < \sigma_2^2 < \sigma'^2_2 < s_2$, and $\sigma_1 = \sigma'_1 < s_1$ (from which we obtain $a_1 = a'_1$ and $\mu_1 = \mu'_1$, (32)). Considering the previous constraints in (34), and multiplying both sides of (34) by $e^{\sigma_1^2 t^2}$ and taking limit as t goes to infinity, we obtain the necessary condition $\pi(1, 1) = \pi'(1, 1)$ (hence terms in $e^{-\sigma_1^2 t^2}$ disappear from (34)). Then multiplying the two sides of (34) by $e^{s_1 t^2}$ and taking the limit as t goes to infinity, we necessarily obtain $\pi(2, 1) = \pi'(2, 1)$ (hence terms in $e^{-s_1^2 t^2}$ disappear from (34)). Finally, multiplying both sides of (34) (with only two terms at this stage) by $e^{\sigma_2^2 t^2}$ and taking the limit as t goes to infinity, we necessarily obtain $\pi(2, 2) = 0$, which is absurd. In any situation such that $(\sigma_1^2, \sigma_2^2) \neq (\sigma'^2_1, \sigma'^2_2)$ the same technique is applied, always leading to an absurd conclusion. In this way it is necessarily established that $(\sigma_1^2, \sigma_2^2) = (\sigma'^2_1, \sigma'^2_2)$. From this remark and (32), we obtain that $a_1 = a'_1$, $a_2 = a'_2$, $\mu_1 = \mu'_1$ and $\mu_2 = \mu'_2$. Including the various identifications thus obtained in (34), we have:

$$\begin{aligned} 0 &= \pi(1)(\pi(1, 1) - \pi'(1, 1)) f_{(m_1, s_1)}(z_1) e^{it(a_1 z_1 + \mu_1) - \sigma_1^2 t^2} \\ &\quad + \pi(1)(\pi(1, 2) - \pi'(1, 2)) f_{(m_1, s_1)}(z_1) e^{itm_2 - s_2 t^2} \\ &\quad + \pi(2)(\pi(2, 1) - \pi'(2, 1)) f_{(m_2, s_2)}(z_1) e^{itm_1 - s_1 t^2} \\ &\quad + \pi(2)(\pi(2, 2) - \pi'(2, 2)) f_{(m_2, s_2)}(z_1) e^{it(a_2 z_1 + \mu_2) - \sigma_2^2 t^2}. \end{aligned} \tag{35}$$

The right-hand side of (35) being a linear combination of linearly independent functions, we obtain that $\pi(i, j) = \pi'(i, j)$, for all i and j in $\{1, 2\}$, which concludes the proof for this first case. The other cases $\sigma_1^2 = \sigma_2^2$ (and $a_1 < a_2$) or $a_1 = a_2$ (and $\sigma_1^2 < \sigma_2^2$) are solved in the same way by using (32)–(35).

Remarks. (i) If we assume $s_1 < s_2$ (which is reasonable in practice), the parameter $\vartheta = (\alpha, \beta, \mu_1, \mu_2, a_1, a_2, \sigma_1^2, \sigma_2^2)$ should be supposed to belong to a compact set $[\delta, 1-\delta]^2 \times [-M, M]^2 \times \mathcal{S}$, where $\mathcal{S}$ is any compact subset of $[0, 1-\delta]^2 \times [\delta, V]^2$, $0 < \delta < 1$ denotes an arbitrary small positive value, and $0 < M < \infty$, $0 < V < \infty$ are arbitrary positive bounds, such that

$$\forall (a_1, a_2, \sigma_1^2, \sigma_2^2) \in \mathcal{S}, \qquad \frac{\sigma_1^2}{1 - a_1^2} < \frac{\sigma_2^2}{1 - a_2^2}. \tag{36}$$

(ii) The previous result can be extended with some extra work to cases corresponding to $K \geqslant 2$ (the previous technique does not use the fact that $\pi(1, 1) = 1 - \pi(1, 2)$ or $\pi(2, 1) = 1 - \pi(2, 2)$ when $K = 2$), but the set $\mathcal{S}$ then becomes much trickier to build.

(iii) Finally, it is worth observing that the case $a_i = 0$ for some values of i in $\{1, \ldots, K\}$, is compatible with this identifiability approach. Thus Markov mixtures of AR(1) processes and sequences of i.i.d. Gaussian random variables, according to expression (4), lead to an identifiable model.

(iv) If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and $a_1 = a_2 = a$, which leads to $s_1 = s_2$, the model is still identifiable. In fact the same kind of proof can be employed using the ordering $\mu_1 < \mu_2$.

Regularity and integrability. We now check essentially that Conditions C3–C6 are satisfied. In order to simplify the expressions, and without loss of generality, we consider $\sigma_1^2 = \sigma_2^2 = 1$ and $a_1 = a_2 = a$ (which corresponds to Remark (iv) above). Let us denote by $\vartheta = (\alpha, \beta, \mu_1, \mu_2, a) = (\vartheta_1, \ldots, \vartheta_5)$. Let us write the two-dimensional likelihood for this parametrization:

$$p_\vartheta(z_1, z_2) = \frac{(1-\alpha)\beta}{\alpha+\beta} T_1(z_1, z_2; \vartheta) + \frac{\alpha\beta}{\alpha+\beta} T_2(z_1, z_2; \vartheta) + \frac{\alpha\beta}{\alpha+\beta} T_3(z_1, z_2; \vartheta) + \frac{\alpha(1-\beta)}{\alpha+\beta} T_4(z_1, z_2; \vartheta)$$

(see (8), with $n = 2$) where

$$T_1(z_1, z_2; \vartheta) = f_{(\mu_1, 1/(1-a^2))}(z_1) f_{(az_1+\mu_1, 1/(1-a^2))}(z_2),$$

$$T_2(z_1, z_2; \vartheta) = f_{(\mu_1, 1/(1-a^2))}(z_1) f_{(\mu_2/(1-a), 1/(1-a^2))}(z_2),$$

$$T_3(z_1, z_2; \vartheta) = f_{(\mu_2/(1-a), 1/(1-a^2))}(z_1) f_{(\mu_1/(1-a), 1/(1-a^2))}(z_2),$$

$$T_4(z_1, z_2; \vartheta) = f_{(\mu_2/(1-a), 1/(1-a^2))}(z_1) f_{(az_1+\mu_2, a^2)}(z_2).$$

Concerning Condition C3, the uniform P_0-integrability of the family $\{\log p_\vartheta(z_1, z_2);\ \vartheta \in \Theta\}$ it is enough to notice that for all $(z_1, z_2) \in \mathbb{R}^2$, and all $\vartheta \in \Theta$, we have

$$\frac{\alpha\beta}{\alpha+\beta} T_2(z_1, z_2; \vartheta) \leqslant p_\vartheta(z_1, z_2) \leqslant 4 \max_{z \in \mathbb{R}} f^2_{(0, 1/(1-a^2))}(z),$$

hence

$$\log\left(\frac{\delta^2(1-\delta/2))}{2\pi}\right) - \left[\frac{1}{2}(z_1^2 + z_2^2) + \frac{M}{\delta}(|z_1| + |z_2|) + \left(\frac{M}{\delta}\right)^2\right] \leqslant \log p_\vartheta(z_1, z_2) \leqslant \log \frac{2}{\pi}.$$

The two sides of the previous inequality being independent of ϑ and P_0-integrable, we thus obtain the desired result. Condition C4 is easy to prove.

Let us now recall that for all $i, j = 1, \ldots, 5$, the expressions for the first- and second-order partial derivatives are given by:

$$\frac{\partial}{\partial\vartheta_i} \log p_\vartheta(z_1, z_2) = \frac{\frac{\partial}{\partial\vartheta_i} p_\vartheta(z_1, z_2)}{p_\vartheta(z_1, z_2)},$$

$$\frac{\partial^2}{\partial\vartheta_i \partial\vartheta_j} \log p_\vartheta(z_1, z_2) = \frac{\frac{\partial^2}{\partial\vartheta_i \partial\vartheta_j} p_\vartheta(z_1, z_2) p_\vartheta(z_1, z_2) - \frac{\partial}{\partial\vartheta_i} p_\vartheta(z_1, z_2) \frac{\partial}{\partial\vartheta_j} p_\vartheta(z_1, z_2)}{(p_\vartheta(z_1, z_2))^2},$$

where, for $\vartheta_1 = \alpha$, $\vartheta_3 = \mu_1$ (the same calculation holding for $\vartheta_2 = \beta$, $\vartheta_4 = \mu_2$), and $\vartheta_5 = a$, we obtain

$$\begin{aligned}
\frac{\partial}{\partial \alpha} p_\vartheta(z_1, z_2) &= \frac{1}{(\alpha+\beta)^2}[-\beta(2\alpha+\beta)T_1(z_1, z_2; \vartheta) + \beta^2(T_2(z_1, z_2; \vartheta) + T_3(z_1, z_2; \vartheta)) \\
&\quad + (1-\beta)\beta^2 T_4(z_1, z_2; \vartheta)], \\
\frac{\partial}{\partial \mu_1} p_\vartheta(z_1, z_2) &= \frac{(1-\alpha)\beta}{\alpha+\beta}\left[\left(z_1 - \frac{\mu_1}{(1-a)}\right)\frac{1-a^2}{1-a} + z_2 - az_1 - \mu_1\right] T_1(z_1, z_2; \vartheta) \\
&\quad + \frac{\alpha\beta}{\alpha+\beta}\left[\left(z_1 - \frac{\mu_1}{(1-a)}\right)\frac{1-a^2}{1-a}\right](T_2(z_1, z_2; \vartheta) + T_3(z_1, z_2; \vartheta)),
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial a} p_\vartheta(z_1, z_2) &= \frac{(1-\alpha)\beta}{\alpha+\beta}\left[\frac{-2a}{1-a^2} + a\left(z_1 - \frac{\mu_1}{1-a}\right)^2\right. \\
&\quad \left. + \left(z_1 - \frac{\mu_1}{1-a}\right)\frac{\mu_1}{(1-a)^2}(1-a^2) + z_1(z_2 - az_1 - \mu_1)\right] T_1(z_1, z_2; \vartheta) \\
&\quad + \frac{\alpha\beta}{\alpha+\beta}\left[\frac{-2a}{1-a^2} - \frac{1-a^2}{(1-a)^2}\left(\mu_1\left(z_1 - \frac{\mu_1}{1-a}\right) + \mu_2\left(z_2 - \frac{\mu_2}{1-a}\right)\right)\right. \\
&\quad \left. + a\left(\left(\frac{z_1-\mu_1}{1-a}\right)^2 + \left(\frac{z_2-\mu_2}{1-a}\right)^2\right)\right] T_2(z_1, z_2; \vartheta) \\
&\quad + \frac{\alpha\beta}{\alpha+\beta}\left[\frac{-2a}{1-a^2} - \frac{1-a^2}{(1-a)^2}\left(\mu_1\left(z_1 - \frac{\mu_2}{1-a}\right) + \mu_2\left(z_2 - \frac{\mu_1}{1-a}\right)\right)\right. \\
&\quad \left. + a\left(\left(\frac{z_1-\mu_2}{1-a}\right)^2 + \left(\frac{z_2-\mu_1}{1-a}\right)^2\right)\right] T_3(z_1, z_2; \vartheta) \\
&\quad + \frac{\alpha(1-\beta)}{\alpha+\beta}\left[\frac{-2a}{1-a^2} + a\left(z_1 - \frac{\mu_2}{1-a}\right)^2\right. \\
&\quad \left. + \left(z_1 - \frac{\mu_2}{1-a}\right)\frac{\mu_2}{(1-a)^2}(1-a^2) + z_1(z_2 - az_1 - \mu_2)\right] T_4(z_1, z_2; \vartheta).
\end{aligned}$$

For the sake of simplicity we do not calculate here the second-order partial derivatives, but from these calculations it can be shown that the absolute values of the two-dimensional likelihood partial derivatives of order 1 and 2 are always dominated by a bivariate function taking the form

$$\mathrm{Pol}_{\vartheta}^{4}(|z_1|, |z_2|)\left(\sum_{i=1}^{4} T_i(z_1, z_2; \vartheta)\right),$$

where $\mathrm{Pol}_{\vartheta}^{4}(\cdot, \cdot)$ is a bivariate polynomial of order 4 whose coefficients depend on ϑ and are uniformly bounded over Θ. On the other hand, for all $(z_1, z_2) \in E^2$, $p_{\vartheta}(z_1, z_2) \geqslant \delta^2 \sum_{i=1}^{4} T_i(z_1, z_2; \vartheta)$. In conclusion the partial derivatives of order 1 and 2 of the two-dimensional log-likelihood function (with respect to the various components of ϑ) are dominated by a bivariate polynomial of order 4 which is $p_{\vartheta_0}(\cdot, \cdot)$-integrable. Finally, we observe that the MSDLE for a mixture of two Gaussian linear AR(1) models is easy to implement since the gradient function of the log-likelihood is analytically known, and classical optimization procedures can be employed to solve $\dot{\ell}_{\vartheta}(z_1^{2k}) = 0$ over Θ, using various initialization conditions.

6. Applications

The goal of this section is to present applications of H4Ms in the fields of neurophysiology (epileptic electroencephalogram signals, and alpha and theta waves), and kinetics (single ion channel analysis). We motivate each application by comparing the expectations of the specialists with H4M properties (i)–(v) described in Section 1, and give further references.

6.1. Epileptic EEG signals

Among the many types of electrical activity in the brain, epileptic electroencephalogram (EEG) signals remain one of the most misunderstood. Various authors have proposed different kinds of model, stochastic and dynamic, to capture the huge complexity of epileptic EEG data series. For a first reading on this subject see, for example, Sackellares *et al.* (2000), Franaszczuk and Bergey (1999), Bergey and Franaszczuk (2001), and references therein. These papers analyse the behaviour of epileptic EEGs and present two different modelling approaches: one based on nonlinear chaotic models, and the other on simple linear models. They conclude that epileptic EEG modelling is an extraordinarily difficult problem which remains open (each method having its advantages and drawbacks). We give a brief account of some of the fundamentals of epileptic EEG signals.

All cerebral activity detectable by EEG is a reflection of synchronous neuronal activity, a state considered normal. Epileptic seizures, however, are abnormal, temporary manifestations of dramatically increased neuronal synchrony, occurring either regionally (partial seizures) or bilaterally (generalized seizure) in the brain. During periods between seizures (called *interictal*) the EEG pattern is of low to medium voltage, irregular and arrhythmic, in contrasts to the organized, rythmic, and self-sustained characteristics of EEG patterns during periods of paroxysmal electrical discharges (ictal). Iasemidis and Sackellarres (1991) study a refinement of the states, by considering the repetitive process of dynamical transitions from the interictal via the pre-ictal (prior to seizure) and ictal and to the post-ictal state (after seizure). Bergey and Franaszczuk (2001) show that the changes occuring at the beginning of a seizure (onset) have not been studied because of the rapidly changing nature of the signal. One of the problems inherent in applying standard signal analysis methods is that most linear and nonlinear methods require long periods of relatively stationary activity. From this description of epilectic EEG patterns, one can propose an H4M model taking into account these main characteristics. Formally, the Markov chain U should have a state space $\mathcal{U}$ with five states, representing the interictal, pre-ictal, onset, interictal, and post-ictal regimes, and a highly structured matrix transition (with a small probability of remaining in the onset state), to reflect the possibility of switching possibilities between these states, and the mixed Markov processes should be autoregressive processes (see Franaszczuk and Bergey 1999, for interictal state modelling with autoregressive processes), with calibrated coefficients (scale and location parameter of the noise, and coefficients of the regression from the past).

6.2. Alpha and theta waves

Waves analysis and classification are crucial in neurophysiology, since they reflect the normality or not of brain activity. Waves of frequency 7.5 Hz and higher are a normal occurrence in the EEG of an awake adult. Lower-frequency waves are classified as abnormal for an awake adult, although they can normally be seen in children or in adults who are asleep. In certain situations, waveforms of otherwise appropriate frequency are considered abnormal because they occur at an inappropriate location or demonstrate irregularities in rhythmicity or amplitude. Some waves are recognized by their shape, head distribution, and symmetry. As a result EEG signals are divided into two groups according to their frequency context and morphomogy characteristics. Novák *et al.* (2001) present an exhaustive classification (with illustrations) of the existing wave forms. We focus our attention on alpha and theta waves, which switch respectively between three and two frequency levels with short and long stationary stages. According to Novák *et al.* (2001), HMM modelling for the alpha wave with a three-state Markov chain seems reasonable (see their Figure 8), while H4M modelling seems much more appropriate for the theta wave because of the abrupt changes and the various piecewise stationary patterns, with trend and notable phase-type feedback effects (see their Figure 10). A very similar sample path is simulated in Section 6.4 below, using a two-state H4M.

6.3. Single ion channel analysis

Ion channels catalyse the diffusion of ions through a membrane into electrical currents of the order of picoamperes (10^{-12} A). The recording of single channel currents shows current levels corresponding to the closed and open state, respectively. Transitions between these two states are very fast and of order of fractions of a millisecond, and appear in the recording as rectangular jumps from one level to the other. Normally, the channels stay open for only a fraction of a second, allowing the flux of tens of thousands of ions through the pore. HMMs provide an efficient approach to the analysis of single channel currents. In fact different states of current levels are supposed during the 'open' state, and are considered as hidden by the noise due to the recording instruments. It is particularly useful for records where the signal-to-noise ratio is low or the channel kinetics is rapid; see Chung *et al.* (1990), Fredkin and Rice (1992), Chung and Gage (1998), see also an excellent overview of the HMM approach to single channel analysis in Quin *et al.* (2000a; 2000b).

Quin *et al.* (2000b) proposed to model the background noise by an autoregressive process, under the strong assumption that the noise depends only on the current state. Venkataramanan and Sigworth (2002) also model the noise as an autoregressive process but make use of a more general description of state-dependent noise. The use of H4Ms in this context should be motivated by regarding the *global signal* (current flux plus noise) as a Markov process (current state discretization is no longer needed), and by considering the open/close mechanism of the pore as a Markovian censoring process of the global signal, following exactly the principle described in formula (4).

6.4. Sample path simulation

In this subsection we show, by considering two-state HMMs and H4Ms, chosen with the same marginal distribution and same switching source, the morphological pattern

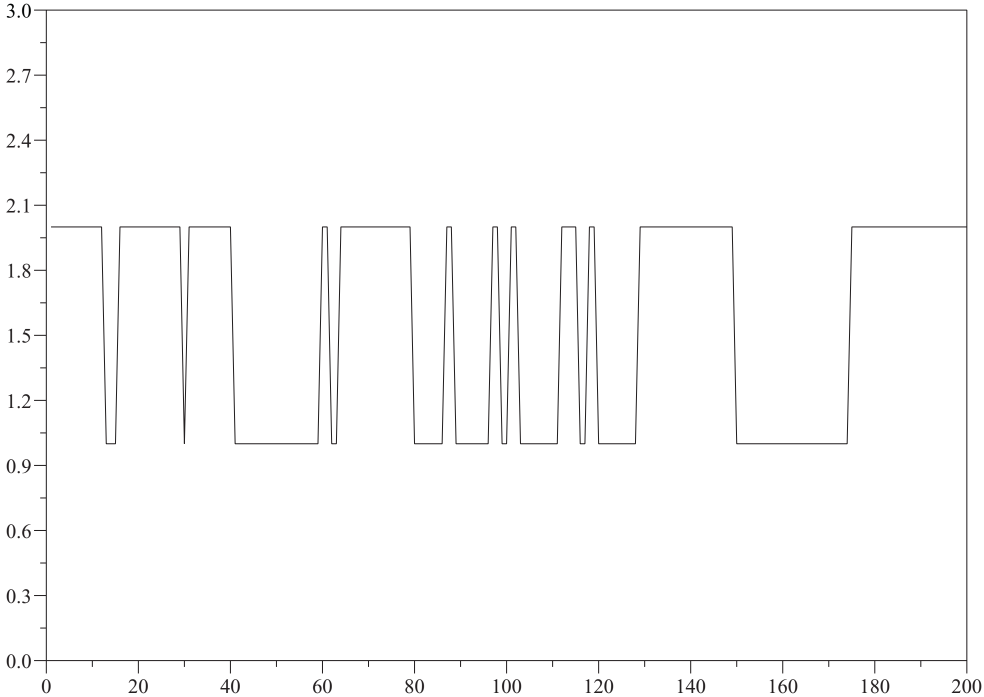

Figure 1. Sample path simulation of the underlying Markov chain U.

differences one can obtain. The models considered have the same underlying chain U, with transition matrix

$$\Pi = \begin{pmatrix} 0.92 & 0.08 \\ 0.08 & 0.92 \end{pmatrix}. \tag{37}$$

For the HMM the conditional law with respect to state 1 is $\mathcal{N}(0, 1)$, and the conditional law with respect to state 2 is $\mathcal{N}(\mu_2, 1)$, with $\mu_2 = 1.5$. For the H4Ms, the AR(1) process X has $a_1 = a$ and noise distribution equal to $\mathcal{N}(0, 1 - a^2)$, while Y has $a_2 = a$ and noise distribution equal to $\mathcal{N}(\mu_2(1 - a), 1 - a^2)$. By construction the HMM and H4Ms previously defined have the same marginal distribution. The following figures, show an HMM sample path (Figure 2), and H4M sample paths, of length $n = 200$, with different choice of a (Figures 3 and 4), using the same chain U (shown in Figure 1).

We observe that in Figure 2 the HMM pattern is very noisy but the global switching scheme is almost clear. In Figure 3 the abrupt changes and the small variance of the jumps from each AR source make the switch design clearer. In Figure 4 the observed sample path is much more difficult to interpret, since the concatened locally stationary sequences are not different enough (because of the importance of the jumps, and the history of each AR

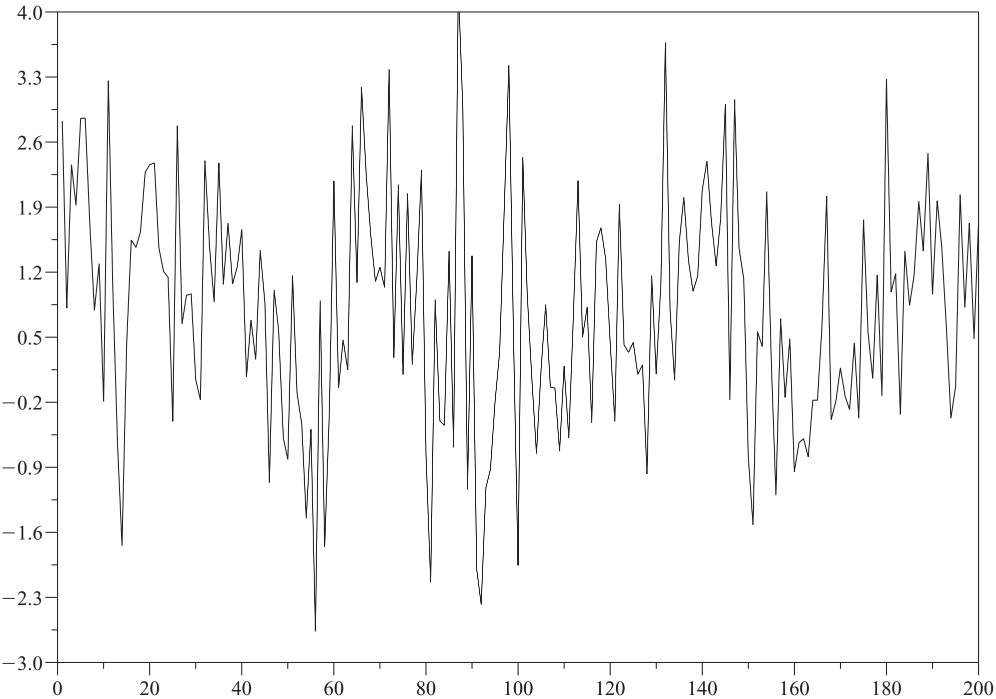

Figure 2. Sample path simulation of an HMM, $\mu_1 = 0$, $\mu_2 = 1.5$, $\sigma_i^2 = 1$, for $i = 1, 2$.

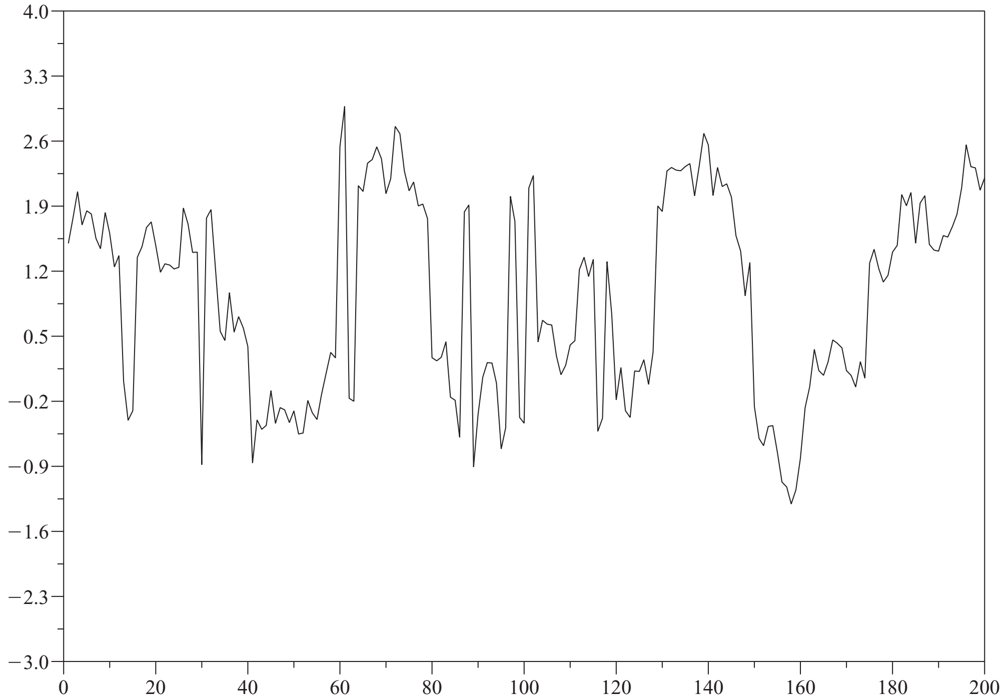

Figure 3. Sample path simulation of an H4M with two AR(1) sources, $a_i = 0.9$ (weakly mixing case), $i = 1, 2$.

source) to detect clearly the instants corresponding to changes of regime. This situation is more ambiguous, in some sense, than in Figure 2, because the resulting process looks like a self-sustained process, where important jumps sometimes occur, and does not have the well-known morphology of a noisy state-space model. Finally, let us remark that pattern shown in Figure 3 imitates quite well the theta wave pattern given in Novák *et al.* (2001).

7. Conclusion

In this paper we have introduced a new missing-data model, the hidden Markov mixture of Markov model (H4M), whose observations come from different independent Markov sources, selection among which at time n is done randomly according to a discrete Markov chain U_n. We observed that such a process is not Markovian, differs clearly from other mixture of Markov Models, and does not belong to the class of hidden Markov models (successive observations are not independent conditionally on a finite-length past of Us). We have proved under mild conditions that the MSDLE, proposed by Rydén and adapted to

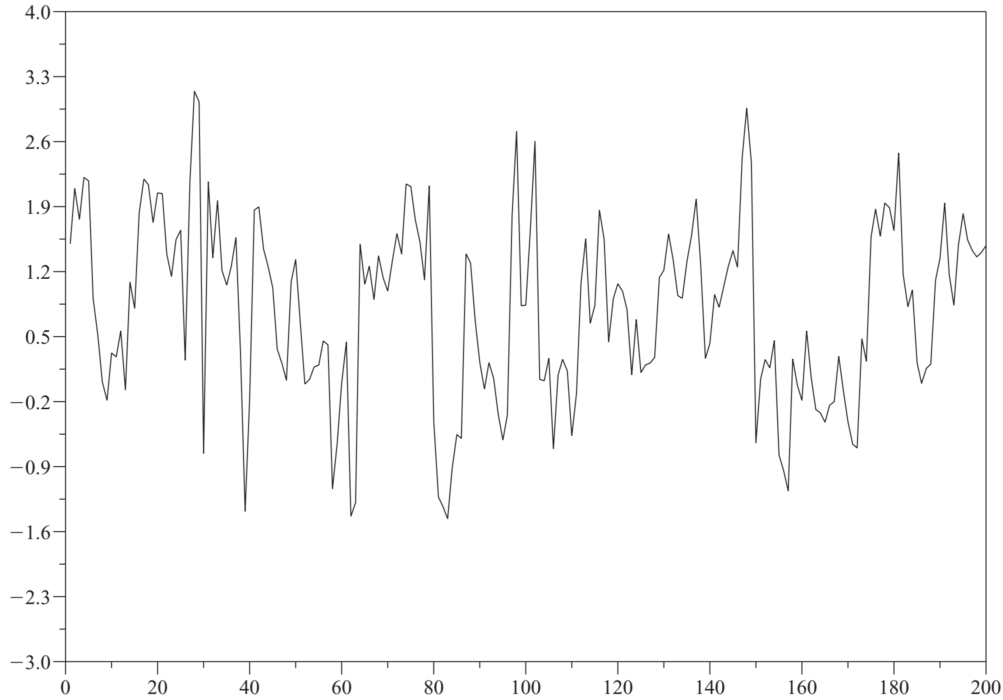

Figure 4. Sample path simulation of an H4M with two AR(1) sources, $a_i = 0.7$ (medium mixing case), $i = 1, 2$.

our case, is consistent and asymptotically distributed. But we have also pointed out that identifiability and the analytic form of the invariant probability densities are in general impossible to derive. To partially address the second difficulty, we proposed a Monte Carlo approach to estimate the split data likelihood when the parametrization of the invariant probability densities is not explicit. However, we exhibit one class of models, the hidden Markov mixture of K linear autoregressive processes of order 1, $K \geqslant 2$, with Gaussian noise, for which all the conditions needed for $\sqrt{n}$-consistency of the MSDLE are satisfied, except for the classical singularity of the covariance matrix involved in the asymptotic normality result. Finally, it seems that H4Ms may be useful models in areas such as neurophysiology and kinetics, which deal with data series with abrupt changes, and locally stationary sequences.

There is scope for future extension of this preliminary work in two directions: (i) research into simple conditions on systems of dynamic equations and families of noise distributions ensuring identifiability of certain classes of H4Ms; and (ii) the study of the very challenging *exact* maximum likelihood estimator, where we claim that parametrization of invariant probability densities is less crucial (in the case of uniform exponential memorylessness of the initial conditions on the $X^{[i]}$, for example).

Acknowledgements

The author would like to thank Laurence Denat for encouragement and insightful comments. He is also grateful to the referees for constructive comments that led to a clearer presentation, and more complete analysis of the problem.

References

Bakry, D., Milhaud, X. and Vandekerkhove, P. (1997) Statistique de chaînes de Markov cachées à espace d'états fini. Le cas non stationnaire. *C. R. Acad. Sci. Paris Sér. I Math.*, **325**, 203–206.

Bar-Shalom, Y. and Li, X.R. (1993) *Estimation and Tracking: Principles, Techniques, and Software.* Norwood, MA: Artech House.

Baraud, Y., Comte, F. and Viennet, G. (2001) Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.*, **29**, 839–875.

Baum, L.E. and Petrie, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **37**, 1554–1563.

Benesch, T. (2001) The Baum–Welch algorithm for parameter estimation of Gaussian autoregressive mixture models. *J. Math. Sci. (New York)*, **105**, 2515–2518.

Bergey, G.K. and Franaszczuk, P.J. (2001) Epileptic seizures are characterized by changing signal complexity. *Clinical Neurophysiology*, **112**, 241–249.

Bickel, P.J., Ritov, Y. and Rydén, T. (1998) Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26**, 1614–1635.

Billingsley, P. (1995) *Probability and Measure*, 3rd edition. Chichester: Wiley.

Cai, J. (1994) A Markov unconditional variance in ARCH. *J. Business Econom. Statist.*, **12**, 309–316.

Chan, K.S. and Tong, H. (1998) A note on testing for multi-modality with dependent data. Unpublished.

Chauveau, D. and Vandekerkhove, P. (2001) An estimator of the entropy to control the stability of Markovian dynamical systems. Reprint.

Chib, S., Kim, S. and Shepard, N. (1998) Stochastic volatility: Likelihood inference and comparison with ARCH models. *Rev. Econom. Stud.*, **65**, 361–394.

Chung, S.H. and Gage, P.W. (1998) Signal processing techniques for channel current analysis based on hidden Markov models. *Methods in Enzymology*, **293**, 420–438.

Chung, S.H., Moore, J., Xia, L., Premkumar, L.S. and Gage, P.W. (1990) Characterization of single channel currents using digital signal processing techniques based on hidden Markov models. *R. Soc. Lond. Philos. Trans. Ser. B*, **329**, 265–285.

Dacunha-Castelle, D. and Duflo, M. (1993) *Probabilités et Statistiques, 2. Problèmes à Temps Mobile.* Paris: Masson.

de Finetti, B. (1959) La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista. In Centro Internazionale Matematico Estivo, *Induzione e Statistica*, pp. 1–115. Rome: Istituto Matematico dell Università.

Dégerine, S. and Zaïdi, A. (2002) Separation of an intantaneous mixture of Gaussian autoregressive sources by the exact maximum likelihood approach. *Submitted to IEEE Trans. Signal Process.*

Diaconis, P. and Freedman, D. (1980) De Finetti's theorem for Markov chains. *Ann. Probab.*, **8**, 115–130.

Douc, R. and Matias, C. (2001) Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, **7**, 381–420.

Douc, R., Moulines, E. and Rydén, T. (2004) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, **32**(5).

Doukhan, P. (1994) *Mixing: Properties and Examples*, Lecture Notes in Statist. 85. New York: Springer-Verlag.

Fortini, S., Ladelli, L., Petris, G. and Regazzini, E. (2002) On mixtures of distributions of Markov chains. *Stochastic Process Appl.*, **100**, 147–165.

Franaszczuk, P.J. and Bergey, G.K. (1999) An autoregressive method for the measurement of synchronization of interictal and ictal EEG signals. *Biol. Cybernet.*, **81**, 3–9.

Francq, C. and Roussignol, M. (1998) Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum likelihood estimator. *Statistics*, **32**, 151–173.

Fredkin, D.R. and Rice, J.A. (1987) Correlation functions of a function of finite-state Markov process with application to chanel kinetics. *Math. Biosci.*, **87**, 161–172.

Fredkin, D.R. and Rice, J.A. (1992) Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Roy. Soc. Lond. Ser. B*, **249**, 125–132.

Freedman, D. (1962) Mixture of Markov processes. *Ann. Math. Statist.*, **33**, 114–118.

Garcia, R. and Perron, P. (1996) An analysis of the real interest rate under regime shift. *Rev. Econom. Statist.*

Hamilton, J.D. (1989) A new approach to the economic analysis of non-stationary time series and the business cycle. *Econometrica*, **57**, 357–384.

Hamilton, J.D. and Susmel, R. (1994) Autoregressive conditional heteroskedasticity and changes in regime. *J. Econometrics*, **64**, 307–333.

Iasemidis, L.D. and Sackellares, J.C. (1991) The evolution with time of the spatial distribution of the largest Lyapounov exponent on the human epileptic cortex. In D. Duke and W. Pritchard (eds), *Measuring Chaos in the Human Brain*. Singapore: World Scientific.

Jalali, A. and Pemberton, J. (1995) Mixture models for time series. *J. Appl. Probab.*, **32**, 123–138.

Ji, C., Snapp, R. and Psaltis, D. (1990) Generalizing smoothness constraints from discrete samples. *Neural Comput.*, **2**(2), 188–197.

Juang, B.H. and Rabiner, L.R. (1991) Hidden Markov models for speech recognition. *Technometrics*, **33**, 251–272.

Krishnamurthy, V. and Rydén, T. (1998) Consistent estimation of linear and non-linear autoregressive models with Markov regime. *J. Times Ser. Anal.*, **19**, 291–307.

LeGland, F. and Mevel, L. (2000) Exponential forgetting and geometric ergodicity in Hidden Markov Models. *Math. Control Signals Systems*, **13**(1), 63–93.

Leroux, B.G. (1992) Maximum likelihood estimation for Hidden Markov models. *Stochastic Process. Appl.*, **20,** 545–558.

Neuts, M.F. (1994) *Matrix-Geometric Solutions in Stochastic Models. An Algorithmic Approach.* Baltimore, MD: Johns Hopkins University Press.

Novák, D., Lhotská, L., Eck, V. and Sorf, M. (2001) EEG and VEP signal processing. Preprint, Czech Technical University.

Pham, D.-T. and Garat, P. (1997) Blind separation of mixture of independent sources through a quasi-maximum likelihood approach, *IEEE Trans. Signal Process.*, **4**, 1712–1725.

Quin, F., Auerbach, A. and Sachs, F. (2000a) A direct optimization approach to hidden Markov modelling for single channel kinetics. *Biophys. J.*, **79**, 1915–1927.

Quin, F., Auerbach, A. and Sachs, F. (2000b) Hidden Markov modelling for single channel kinetics with filtering and correlated noise. *Biophys. J.*, **79**, 1928–1944.

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–284.

Rydén, T. (1994) Consistent and asymptotically normal parameter estimates for hidden Markov models. *Ann. Statist.*, **22**, 1884–1895.

Sackellares, J.C., Iasemidis, L.D., Shiau, D.-S., Gilmore, R. and Roper, S.N. (2000) Epilepsy – when chaos fails. In K. Lehnertz, J. Arnhold, P. Grassberger and C.E. Elger (eds), *Chaos in the Brain?* Singapore: World Scientific.

Stout, W. (1974) *Almost Sure Convergence.* New York: Academic Press.

Teicher, H. (1963) Identifiability of finite mixture. *Ann. Math. Statist.*, **34**, 1265–1269.

Tong, H. (1990) *Non-linear Time Series.* New York: Oxford University Press.

Tugnait, J.K. (1982) Detection and estimation for abruptly changing systems. *Automatica*, **18**, 607–615.

Venkataramanan, L. and Sigworth, F.J. (2002) Applying hidden Markov models to the analysis of single ion channel activity. *Biophys. J.*, **82**, 1930–1942.

Wong, C.S. and Li, W.K. (2000) On a mixture autoregressive model. *J. R. Statist. Soc., Ser. B,* **62**, 95–115.

Wong, C.S. and Li, W.K. (2001) On a logistic mixture autoregressive model. *Biometrika*, **88**, 833–846.

Zhengyan, L. and Chuanrong, L. (1996) *Limit Theory for Mixing Dependent Random Variables*. Dordrecht: Kluwer Academic Publishers.

Received October 2003 and revised May 2004