Signal Processing 90 (2010) 2899–2910

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Unsupervised segmentation of new semi-Markov chains hidden with long dependence noise

Jérôme Lapuyade-Lahorgue, Wojciech Pieczynski *

Institut Telecom, Telecom Sudparis, CITI Department, CNRS UMR 5157, 9, rue Charles Fourier, 91000 Evry, France

ARTICLE INFO

Article history:
Received 29 July 2009
Received in revised form
28 January 2010
Accepted 9 April 2010
Available online 21 April 2010

Keywords:
Hidden semi-Markov chains
Unsupervised signal segmentation
Long dependence noise
Iterative conditional estimation
Unsupervised signal segmentation
Unsupervised image segmentation

ABSTRACT

The hidden Markov chain (HMC) model is a couple of random sequences (X,Y), in which X is an unobservable Markov chain, and Y is its observable "noisy version". The chain X is a Markov one and the components of Y are independent conditionally on X. Such a model can be extended in two directions: (i) X is a semi-Markov chain and (ii) the distribution of Y conditionally on X is a "long dependence" one. Until now these two extensions have been considered separately and the contribution of this paper is to consider them simultaneously. A new "semi-Markov chain hidden with long dependence noise" model is proposed and it is specified how it can be used to recover X from Y in an unsupervised manner. In addition, a new family of semi-Markov chains is proposed. Its advantages with respect to the classical formulations are the low computer time needed to perform different classical computations and the facility of its parameter estimation. Some experiments showing the interest of this new semi-Markov chain hidden with long dependence noise are also provided.

1. Introduction

The classical hidden Markov chain (HMC) model is a couple of random sequences (X,Y), with $X=(X_1,\ldots,X_N)$ and $Y=(Y_1,\ldots,Y_N)$, in which X is an unobservable Markov chain, and Y is its observable noisy version Throughout this paper, we will assume that each X_n takes its values in $\Omega=\{\omega_1,\ldots,\omega_K\}$, and each Y_n takes its values in R. The aim of the Bayesian segmentation is to estimate the hidden realization x of X from an observed realization y of Y. One possible way of such an estimation is to set, for $n=1,\ldots,N$, $x_n^* = \arg\max_{x_n} p(x_n|y)$. This estimation, which will be used everywhere below, is called the "Marginal Posterior Mode" (MPM). The estimate $x^*=(x_1^*,\ldots x_N^*)$ is computable, even for very large N, when the distribution $p(x,y)$ is not too complex. The most classical model allowing the computation of x^* is the "hidden Markov chain" (HMC) whose distribution $p(x,y)$ is given by

$$p(x,y) = p(x_1)\prod_{n=1}^{N-1} p(x_{n+1}|x_n)\prod_{n=1}^{N} p(y_n|x_n) \tag{1.1}$$

The hidden chain X is then a Markov chain, the random variables Y_n are independent conditionally on X, and they verify $p(y_n|x)=p(y_n|x_n)$. HMCs are widely used and are quite efficient in numerous situations [5,11,15,16,19,24,26,31].

However, both assumptions "Markovianity of X" and "independence of the components of Y conditionally on X" can turn out to be too strong in some situations and two following extensions have been proposed:

(i) The first extension consists in replacing the Markov X by a "semi-Markov", also called the "explicit duration", or "variable-duration", chain. One then obtains a "hidden semi-Markov chain" (HSMC), which still makes it possible to estimate X from Y and which is of interest, with respect to the classical HMC, in numerous situations [1,4,12,13,17,18,20,25,27,31,32,

* Corresponding author.
E-mail address: Wojciech.Pieczynski@it-sudparis.eu (W. Pieczynski).

0165-1684/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2010.04.008

34,35,39]; also see the recent overview [40] with a rich bibliography on the subject;

(ii) In the second extension, recently proposed in [21], the very simple distribution of Y conditional on X is extended to a "long dependence" distribution. Such a model, where X remains a Markov chain, can have two kinds of interpretations. Many phenomena are modeled by long dependence processes [10,33,38] and there can be "changing points", or "switches", in which the model parameters change. Detecting such switches is then an important problem [6,36,37]. The second interpretation corresponds to some X with "physical" existence, as in all the examples dealt with in the context of HMC. Of course, both interpretations are close to each other: switches of parameters can be modelled by a latent variable X.

The aim of this paper, which extends the results contained in [22,23], is to simultaneously consider both extensions. A new "semi-Markov chain hidden with long dependence noise" model is proposed and it is specified how it can be used to recover X from Y in an unsupervised manner. Thus, the model proposed can either be seen as an extension of the model (i) to "long dependence noise", or as an extension of the model (ii) to "semi-Markov" chain X.

In addition, a new family of semi-Markov chains (NSMCs) is proposed. This new family can be used either in the classical context of the semi-Markov chains hidden with "independent noise" (1.1), or in the new context extended to long dependence noise. The interest of using NSMCs lies in low computer time needed to perform different classical computations and in the ease with which its parameters can be estimated.

To avoid any confusion let us recall that there are different research works in which a long dependence process is approximated by a "Markov switching" model, which is also called a "mixture" model [9]. This is not the purpose of this work. In fact, in the model used here $p(y|x)$ is a long correlation distribution for each sequence x, and thus one does not need to approach $p(y|x)$ by any Markov switching distribution.

The paper is organized as follows. The next section is devoted to the presentation of the new semi-Markov chain hidden with the classical independent noise. In Section 3 one recalls the recent Markov chain hidden with long dependence Gaussian noise model [21]. The new "semi-Markov chains hidden with long dependence noise" family of models is introduced in Section 4, while Section 5 is devoted to the related parameter estimation. Different experiments are described in Section 6, and the last Section 7 contains conclusions and perspectives.

2. Hidden semi-Markov chains

2.1. Classical formulation of hidden semi-Markov chains

Let $X=(X_n)_{1\leq n\leq N}$ and $Y=(Y_n)_{1\leq n\leq N}$ be two stochastic chains as above, and let $U=(U_n)_{1\leq n\leq N}$ be a stochastic chain taking its values in $\Lambda=\{0,\ldots,N-1\}$. To obtain a classical semi-Markov distribution for X, which will be called "classical" SMC and denoted by "CSMC", one possible way is to assume that (X,U) is a Markov chain whose distribution is defined by the following. Let δ be the Kronecker function, which is defined by $\delta_a(b)=1$ for $a=b$, and $\delta_a(b)=0$ for $a\neq b$. The distribution of (X,U) is given by $p(x_1,u_1)$ and the transitions $p(x_{n+1},u_{n+1}|x_n,u_n)=p(x_{n+1}|x_n,u_n)p(u_{n+1}|x_n,u_n,x_{n+1})$ verifying

$$p(x_{n+1}|x_n,u_n)=\delta_{x_n}(x_{n+1}) \text{ if } u_n>0, \text{ and } p(x_{n+1}|x_n) \text{ if } u_n=0 \tag{2.1}$$

$$p(u_{n+1}|x_{n+1},x_n,u_n)=\delta_{u_n-1}(u_{n+1}) \text{ if } u_n>0, \text{ and } p(u_{n+1}|x_{n+1}) \text{ if } u_n=0, \tag{2.2}$$

with $p(x_{n+1}|x_n)=0$ if $x_{n+1}=x_n$. Classically u_n is the exact remaining sojourn time in x_n. The advantage of the semi-Markov distributions over the Markov ones is that in the former the distribution of U_n is of any form, while it is necessarily of exponential form in the latter.

Setting

$$p(y|x,u)=\prod_{n=1}^{N}p(y_n|x_n) \tag{2.3}$$

the triplet (X,U,Y) is the classical "hidden semi-Markov chain", which will be denoted by "CHSMC".

Let $V=(X,U)$. As (V,Y) is a classical hidden Markov chain, the classical computation used in HMC can be used to compute $p(v_n|y)=p(x_n,u_n|y)$, which gives

$$p(x_n|y)=\sum_{u_n}p(x_n,u_n|y) \tag{2.4}$$

2.2. New class of hidden semi-Markov chains

As specified above, in the classical HSMC the chain $U=(U_n)_{1\leq n\leq N}$ takes its values from $\Lambda=\{0,\ldots,N-1\}$, where N is the size of the observed sample. This can become a drawback when N is large; for example, in image processing area the sample size, which is an image size, can be about one million. To remedy this, let us consider a particular non-standard semi-Markov chain (X,U) introduced in [22,23].

Eqs. (2.1) and (2.2) are kept, but there are two modifications with respect to the classical definition of the previous sub-section. The first one is to consider $\Lambda=\{0,\ldots,L-1\}$, with a fixed L which no longer depends on N. In practice, L will be taken of small size compared to N. The second modification, which makes the model nonstandard, is to relax the constraint "$p(x_{n+1}|x_n)=0$ if $x_{n+1}=x_n$". Then u_n is the minimum remaining sojourn time in x_n, and not the exact remaining sojourn time as in the classical model. In other words, for $u_n=0$, the random variable X_{n+1} can remain in the same state $X_n=x_n$ with nonnull probability $p(x_{n+1}=x_n|x_n)$.

Finally, the new model, which will be denoted by "NSMC", is defined by a Markov chain $(X,U)=(X_n,U_n)_{1\leq n\leq N}$, where each X_n takes its values from $\Omega=\{\omega_1,\ldots,\omega_K\}$, each U_n takes its values from $\Lambda=\{0,\ldots,L-1\}$, and whose transitions are defined by (2.1) and (2.2), without the hypothesis "$p(x_{n+1}|x_n)=0$ if $x_{n+1}=x_n$".

Let us notice two further points highlighting the position of the NSMC with regard to the CSMC:

(1) A CSMC is a Markov chain when the distribution of the exact sojourn time follows a geometric distribution.

The situation is quite different in NSMCs: an NSMC is a Markov chain when the distribution of the minimal sojourn time is a Dirac mass: $p(u_n=0|x_n)=1$ for each $n=1,\ldots,N$ and each $x_n \in \{\omega_1,\ldots,\omega_K\}$. In this case the transition matrices $p(x_{n+1}|x_n)$ related to the NSMC are the transition matrices related to the Markov chain.

(2) The distribution of an NSMC can be seen as a distribution of a CSMC with the difference that in NSMC the exact sojourn time distribution is not known explicitly. More precisely, let us consider a NSMC defined by the family of distributions $p(u_{n+1}|x_{n+1})$ on $\Lambda=\{0,\ldots,L-1\}$, and the transitions $p(x_{n+1}|x_n)$. Let $p^*(x_{n+1}|x_n)$ be the transitions obtained from the transitions $p(x_{n+1}|x_n)$ by $p^*(x_{n+1}|x_n)=0$ for $x_n=x_{n+1}$, and $p^*(x_{n+1}|x_n)=p(x_{n+1}|x_n)/(1-p(x_{n+1}=x_n|x_n))$ for $x_n\neq x_{n+1}$, and let $p^*(u_{n+1}|x_{n+1})$ be the distributions on $\mathrm{N}=\{0,1,\ldots,N,\ldots\}$ defined by

$$p^*(u_{n+1}=d|x_{n+1}=\omega_j)=p(x_{n+1+d}\neq\omega_j|x_n=\omega_j)\times$$
$$\sum_{k=1}^{d}[\sum_{s_1+\ldots+s_k=d}[p(u_{n+1}=s_1|x_{n+1}=\omega_j)p(u_{n+1+s_1}=s_2|x_{n+1+s_1}=\omega_j)p(u_{n+1+s_1+s_2}=s_3|x_{n+1+s_1+s_2}=\omega_j)\ldots$$
$$p(u_{n+1+s_1+\ldots+s_{k-1}}=s_k|x_{n+1+s_1+\ldots+s_{k-1}}=\omega_j)\times$$
$$p(x_{n+1+s_1}=\omega_j|x_{n+1}=\omega_j)p(x_{n+1+s_1+s_2}=\omega_j|x_{n+1}=\omega_j)\ldots p(x_{n+1+s_1+\ldots+s_{k-1}}=\omega_j|x_{n+1}=\omega_j)]]$$

Then the NSMC distribution given by the distributions $p(u_{n+1}|x_{n+1})$ and the transitions $p(x_{n+1}|x_n)$ is identical to the CSMC distribution given by the distributions $p^*(u_{n+1}|x_{n+1})$ and the transitions $p^*(x_{n+1}|x_n)$. Of course such a distribution is difficult to encircle. However, as far as the problem of Bayesian segmentation is concerned, this does not appear to be a drawback. Moreover, it allows very simple, flexible, and rich parameterization, which is given by the choice of L and a probability distribution, parameterized or not, on $\Lambda=\{0,\ldots,L-1\}$.

To obtain a hidden NSMC, which will be called "NHSMC", one considers the same distribution of Y conditional on (X,U) given by (2.3).

Remark 2.1. Both CHSMC and NHSMC can thus be considered as particular classical HMC (V,Y), with $V=(X,U)$. Such a representation is of interest as all classical computations known in HMC can be used in CHSMC and NHSMC, once the particular form (2.1) and (2.2) of the transitions and the particular form (2.3) of the noise have been taken into account.

Remark 2.2. The CSMCs $V=(X,U)$ have been introduced to extend the Markov chains and thus, in theory, each variable U_n should take its values from the set of natural numbers N. However, as the number of observations N is finite, one can consider that each variable U_n takes its values from $\Lambda=\{0,\ldots,N-1\}$, as specified at the beginning of Section 2. In practice, this set is often restricted to $\Lambda=\{0,\ldots,D\}$, where D is the maximal duration independent of N [39]. As the classical Markov chain does not belong to the family of such models, doing so is equivalent to considering an approximation of the "true" SMC. Now, the new SMC proposed here is quite close to this approximation: the "exact" duration just becomes the "minimal" duration. As specified above, doing so leads to a real semi-Markov chain.

Remark 2.3. In both CHSMCs and NHSMCs (2.3) can be easily extended to $p(y|x,u)=\prod_{n=1}^{N}p(y_n|x_n,u_n)$, where $p(y_n|x_n,u_n)$ varies with u_n. In fact, such an extension does not interfere in different computations of interest and it can be justified in some situations. For example, let us consider a CHSMC and let us imagine that $Y=(Y_1,\ldots,Y_N)$ is a line in a digital image. Let $\Omega=\{\omega_1,\omega_2\}$, where ω_1 is "forest" and ω_2 is "water". If $x_n=\omega_1$ the distribution $p(y_n|x_n=\omega_1,u_n)$, which models the variability of the forest, can depend on u_n. In fact, the aspect of forest can be different in spots bordering on water ($u_n=0$) than elsewhere.

Remark 2.4. As specified above, both CHSMCs and NHSMCs are particular classical HMCs (V,Y), with $V=(X,U)$. Knowing that the classical HMCs can be extended to the "pairwise" Markov chains (PMCs [28]) with a noticeable increase in efficiency in unsupervised data classification [8], let us briefly specify how this extension can be applied to both CHSMCs and NHSMCs. In CHSMCs and NHSMCs one has $p(v_{n+1}, y_{n+1}|v_n,y_n)=p(v_{n+1}|v_n)p(y_{n+1}|v_{n+1})$, with $p(v_{n+1}|v_n)$ defined by (2.1) and (2.2). In a "pairwise" extension it would be of the form $p(v_{n+1},y_{n+1}|v_n,y_n)=p(v_{n+1}|v_n,y_n)p(y_{n+1}|v_{n+1},v_n,y_n)$, where $p(v_{n+1}|v_n,y_n)$ is still given by (2.1) and (2.2), with y_n as a constant, and $p(y_{n+1}|v_{n+1},v_n,y_n)$ is possibly kept in its general form. Let us underline the fact that in such a "pairwise" extension $V=(X,U)$ is no longer necessarily a Markov chain [29], and thus X is no longer necessarily a semi-Markov chain. However, $V=(X,U)$ is Markovian conditionally on Y and thus estimating it from Y remains feasible. Finally, such a model (X,U,Y) is a triplet Markov chain [30], but neither X nor (X,U) is Markovian.

3. Semi-Markov chains hidden with general Gaussian noise (GGN-HSMCs)

3.1. Markov chains hidden with general Gaussian noise (GGN-HMC)

Let $V=(V_n)_{1\leq n\leq N}$ be a Markov chain taking its values from $\Delta=\{\lambda_1,\ldots,\lambda_M\}$, and let $Y=(Y_n)_{1\leq n\leq N}$ be a stochastic real valued sequence. In the next sub-section V will be assumed to be a CSMC or NSMC taking its values from $\Delta=\Omega\times\Lambda$, dealt with in the previous section. However, the

formulas of interest presented in this sub-section are valid for any Markov chain $V=(V_n)_{1\leq n\leq N}$ and thus its general form will be kept for the moment.

For each $n=1,\ldots,N$ we will set $V_1^n=(V_1,\ldots,V_n)$, $Y_1^n=(Y_1,\ldots,Y_n)$, and $v_1^n=(v_1,\ldots,v_n)$, $y_1^n=(y_1,\ldots,y_n)$ for their realizations.

The couple (V,Y) is called a "pairwise partially Markov chain" (PPMC) if its distribution is given by $p(v_1,y_1)$ and the transitions $p(v_{n+1},y_{n+1}|v_1^n,y_1^n)$ verifying

$$p(v_{n+1},y_{n+1}|v_1^n,y_1^n)=p(v_{n+1}|v_n,y_1^n)p(y_{n+1}|v_n,v_{n+1},y_1^n) \tag{3.1}$$

Such a model is thus Markovian with respect to V, but is not necessarily Markovian with respect to Y. One can show [21] that $p(v|y)$ is Markovian with the distribution given by

$$p(v_1|y)=\frac{p(v_1,y_1)\beta_1(v_1)}{\sum_{v_1}p(v_1,y_1)\beta_1(v_1)}; \tag{3.2}$$

$$p(v_{n+1}|v_n,y)=\frac{p(v_{n+1}|v_n,y_1^n)p(y_{n+1}|v_n,v_{n+1},y_1^n)|\beta_{n+1}(v_{n+1})}{\beta_n(v_n)}, \tag{3.3}$$

where the quantities $\beta_1(v_1),\ldots,\beta_N(v_N)$ are computable using the backward recursion

$$\beta_N(v_N)=1,\quad \beta_n(v_n)=\sum_{v_{n+1}}p(v_{n+1}|v_n,y_1^n)p(y_{n+1}|v_n,v_{n+1},y_1^n)\beta_{n+1}(v_{n+1}) \tag{3.4}$$

Thus, $p(v_1|y)$ and $p(v_{n+1}|v_n,y)$ can be computed once the probabilities $p(v_{n+1}|v_n,y_1^n)$ and $p(y_{n+1}|v_n,v_{n+1},y_1^n)$ are known.

Let us consider these probabilities of particular Gaussian form, as introduced in [21]. One assumes that $p(v_{n+1}|v_n,y_1^n)=p(v_{n+1}|v_n)$, and the probabilities $p(y_{n+1}|v_n,v_{n+1},y_1^n)$ are assumed to be defined by M^2 Gaussian distributions on R^N in the following way. For each $(\lambda_i,\lambda_j)\in\Delta^2$, let q^{ij} be a Gaussian distribution on R^N. Thus, for each $n=1,\ldots,N-1$ and each $y_1^n=(y_1,\ldots,y_n)$, one has a Gaussian conditional distribution $q^{ij}(y_{n+1}|y_1^n)$ on R. Then $p(y_{n+1}|v_n,v_{n+1},y_1^n)$ are defined by

$$p(y_{n+1}|v_n=\lambda_i,v_{n+1}=\lambda_j,y_1^n)=q^{ij}(y_{n+1}|y_1^n) \tag{3.5}$$

The important point is that in the Gaussian case considered, for each $(\lambda_i,\lambda_j)\in\Delta^2$ the distributions $q^{ij}(y_1)$, $q^{ij}(y_2|y_1^1),\ldots,q^{ij}(y_{n+1}|y_1^n),\ldots,q^{ij}(y_N|y_1^{N-1})$ are recursively computable in a classical way, each $q^{ij}(y_{n+1}|y_1^n)$ being computable from the previous $q^{ij}(y_n|y_1^{n-1})$.

Such a model will be called a "Markov chain hidden with general Gaussian noise" (GGN-HMC).

Remark 3.1. Let us notice that according to the model, for each $i=1,\ldots,M$ the distribution q^{ii} is the distribution of Y conditional on $v_1=v_2=\cdots=v_N=\lambda_i$. For $i\neq j$, the distributions q^{ij} have no immediate interpretation. However, the model can be simplified by setting $p(y_{n+1}|v_n,v_{n+1},y_1^n)=p(y_{n+1}|v_{n+1},y_1^n)$ and then there are M Gaussian distributions $q^1,\ldots,q^M$ on R^N used, with $p(y_{n+1}|v_{n+1}=\lambda_j,y_1^n)=q^j(y_{n+1}|y_1^n)$. Such a simplified model, in which $q^1,\ldots,q^M$ are easier to interpret, will be considered in experiments below.

Finally, the posterior marginal distribution $p(v_n|y)$ of interest can then be computed in four following recursions [21]:

(i) M^2 forward recursions: for each $i,j=1,\ldots,M$, and for each $n=1,\ldots,N-1$, compute $q^{ij}(y_{n+1}|y_1^n)$ from $q^{ij}(y_n|y_1^{n-1})$;
(ii) compute $\beta_n(v_n)$ for each $n=N,\ldots,1$ by the following backward recursion: $\beta_N(\lambda_i)=1$ for each λ_i, $\beta_n(\lambda_i)=\sum_{\lambda_j}p(v_{n+1}=\lambda_j|v_n=\lambda_i)q^{ij}(y_{n+1}|y_1^n)\beta_{n+1}(\lambda_j)$;
(iii) set $p(v_1|y)=\frac{p(v_1,y_1)\beta_1(v_1)}{\sum_{v_1}p(v_1,y_1)\beta_1(v_1)}$,
$p(v_{n+1}|v_n,y)=\frac{p(v_{n+1}|v_n)p(y_{n+1}|v_n,v_{n+1},y_1^n)\beta_{n+1}(v_{n+1})}{\beta_n(v_n)}$;
(iv) and compute $p(v_n|y)$ for each $n=1,\ldots,N$ by the classical forward recursion: $p(v_1|y)$ given; $p(v_{n+1}|y_1^N)=\sum_{v_n}p(v_n|y_1^N)p(v_{n+1}|v_n,y_1^N)$.
Let us notice that $p(v_n,v_{n+1}|y_1^N)$, which will be needed in the parameter estimation in Section 5, is then given by
(v) $p(v_n,v_{n+1}|y_1^N)=p(v_n|y_1^N)p(v_{n+1}|v_n,y_1^N)$.

Finally, to summarize one can say that in GGN-HMCs the posterior marginal distributions $p(v_n|y_1^N)$ are computable with complexity linear in MN.

3.2. Semi-Markov chains hidden with general Gaussian noise (GGN-HSMCs)

Let us assume now that $V=(X,U)$ is either a CSMC or NSMC, as introduced in Section 2. As V is a Markov chain it is possible to apply the considerations of the previous sub-section resulting in two "Semi-Markov chains hidden with general Gaussian noise" (GGN-HSMCs), which will be denoted by GGN-CHSMCs and GGN-NHSMCs, respectively. In each of them there are two possibilities. The classical one would consist of considering that the distribution of the noise conditional on $V=(X,U)$ only depends on X. The more general model consists of following the Remark 2.3 and considering that this distribution depends on both X and U.

Finally, in all these models $p(v_n|y)$ are computable with reasonable complexity. However, the noise distributions $p(y|v)$ involve a great deal of parameters which could possibly be difficult to determine. For example, taking a GGN-NHSMC in which the noise distribution depends on both X and U, there are $(KL)\times(KL)$ Gaussian distributions on R^N. If these parameters are known, as well as those defining the distribution of the NSMC $V=(X,U)$, it is possible to perform a Bayesian estimation of $V=(X,U)$ from Y. If not, the parameters have to be estimated. It is possible to propose an estimation method once a particular simplified form for the Gaussian distribution had been considered, which is made in the next section.

4. Semi-Markov chains hidden with long dependence Gaussian noise

4.1. General model

Let us consider the following particular Gaussian distributions q^{ij}. First, they are stationary: for each $k=1,\ldots,N-1$ and $n=1,\ldots,N-k$, the marginal densities $q^{ij}(y_n,y_{n+k})$ only depend on k. Second, the Gaussian

distributions $q^{ij}(y_{n+1}|y_1^n)$ only depend on j, which means that $p(y_{n+1}|v_n=\lambda_i,v_{n+1}=\lambda_j,y_1^n)$ are equal to $p(y_{n+1}|v_{n+1}=\lambda_j,y_1^n)=q^j(y_{n+1}|y_1^n)$. This second hypothesis is not essential and what is said below remains valid without it; however, it will be assumed to simplify the model and make it more intuitive. In fact, as specified in Remark 3.1, each q^j is then the distribution of $Y=(Y_n)_{1\le n\le N}$ conditional to $x_1^N=(\underbrace{\omega_j,\ldots,\omega_j}_{N \text{ times}})$.

In other words, each q^j can be seen as a correlated "noise" related to each class ω_j. Taking all q^j stationary, one can imagine that the covariance $\gamma^j(k)$ of $q^j(y_n,y_{n+k})$ decreases with k. One will say that q^j is a "long dependence" distribution if there exist $c^j>0$ and $\alpha^j\in]0,1]$ such that $\gamma^j(k)k^{\alpha^j}\underset{k\to+\infty}{\longrightarrow}c^j$ [33]. The "long dependence" means that the sequence $\gamma^j(k)$ tends to 0 "slowly" enough to imply $\sum_{k=0}^{+\infty}\gamma^j(k)=+\infty$. The long dependence is opposed to the "short dependence" in which $\sum_{k=0}^{+\infty}\gamma^j(k)<+\infty$; in particular, Markov processes are of the "short dependence" kind.

Finally, a stationary GGN-HSMC (V,Y) defined by a semi-Markov chain $V=(X,U)$ and the Gaussian distributions $q^1,\ldots,q^K$ will be called a "long dependence Gaussian noise" HSMC, and will be denoted by LDGN-HSMC, if at least one Gaussian distribution q^j is a long dependence one. According to the nature of the HSMC $V=(X,U)$, which can be of a "new" or "classical" kind, there are two kinds of LDGN-HSMC: the "classical" LDGN-CHSMC and the "new" LDGN-NHSMC.

4.2. Parameterized model

Let us consider the parameterized covariances $\gamma^j(k)=c^j(k+1)^{-\alpha^j}$, with $c^j>0$ and $\alpha^j\in]0,1]$, which will be used in experiments below. There are two corresponding LDGN-NHSMCs (V,Y). In the simplest one the distribution of Y conditional on $V=(X,U)$ only depends on X and thus the model parameters are the following. The distribution of the Markov chain $V=(X,U)$, which takes its values from $\Omega\times\Lambda$ with $\Omega=\{\omega_1,\ldots,\omega_K\}$ and $\Lambda=\{0,\ldots,L-1\}$, is defined by (2.1) and (2.2). In the stationary case considered one needs $p(x_1)$, $p(x_2|x_1)$, and $p(u_1|x_1)$, which involves $K-1+K(K-1)+K(L-1)=K(K+L-1)-1$ parameters. The distribution of Y conditional on X is defined by $3K$ parameters (m^1,c^1,α^1), …, (m^K,c^K,α^K), each (m^j,c^j,α^j) defining q^j. In the extended case the distribution of Y conditional on $V=(X,U)$ is defined by $3KL$ (instead of $3K$) parameters (m^j,c^j,α^j), the distribution of $V=(X,U)$ remaining the same.

Remark 4.1. Concerning the corresponding LDGN-CHSMC case, let us consider the simple case in which the distribution of Y conditional on $V=(X,U)$ only depends on X. As such model is a direct extension of the classical hidden semi-Markov chain, it is of interest to specify what has to be modified in the latter to obtain the former. As the distribution of the semi-Markov chain $V=(X,U)$ is strictly the same in both models, the difference lies in $p(y|x)$. In the first case it is of the form $p(y|x)=\prod_{n=1}^N p(y_n|x_n)$, while it is of the form $p(y|x)=p(y_1|x_1)\prod_{n=2}^N p(y_n|x_n,y_1^{n-1})$. Thus, everything is the same except the fact that $p(y_n|x_n)$ has to be replaced by $p(y_n|x_n,y_1^{n-1})$. The former distribution is very simple while the latter one has to be computed from the Gaussian distributions q^j. In particular, this is the only change needed in the computer programs relating to the classical hidden semi-Markov chains.

However, let us underline the fact that in spite of the simplicity of $p(y_n|x_n,y_1^{n-1})$, which are the same in both LDGN-CHSMC and LDGN-NHSMC cases, the distribution $p(y|x)$ is somewhat complicated. In particular, for each $n=1,\ldots,N$, the Gaussian distribution $p(y_n|x_1^n)$ does depend on all $x_1,\ldots,x_n$ [21].

4.3. Bounded length of memory

When N is large the recursive computation of M^2 (with $M=K$ if the distribution of Y conditional on $V=(X,U)$ only depends on X, and $M=KL$ if not) sequences of the conditional distributions $q^j(y_2|y_1^1)$, $q^j(y_3|y_1^2),\ldots,$ $q^j(y_N|y_1^{N-1})$ can pose computational problems. In fact, the problem comes from the large memory needed to store the matrices defining these distributions. To remedy this, one possible way is to limit the dependence size to a given fixed S and to consider $q^j(y_n|y_{n-S}^{n-1})$ instead of $q^j(y_n|y_1^{n-1})$. Further simplification, which will be used in experiments below, consists of only considering in $q^j(y_n|y_{n-s}^{n-1})$ the observations $y_{n-1},y_{n-2},\ldots,y_{n-k}$ such that $x_n=x_{n-1}=\cdots=x_{n-k}$, and $x_n\neq x_{n-k-1}$. This can be modeled by an auxiliary chain W taking its values from $S=\{0,\ldots,S\}$ and, at each $n=1,\ldots,N$, the variable W_n designates the number $k\le S$ of previous indices $n-1,\ldots,n-k$ such that $x_n=x_{n-1}=\cdots=x_{n-k}$, and $x_n\neq x_{n-k-1}$. Therefore one can say, for $W_n<S$, that W_n is the exact past sojourn time in x_n. The triplet $V=(X,W,U)$ is thus a Markov chain, and (V,Y) is a partially Markov chain whose distribution is given by $p(v_1,y_1)=p(x_1)p(y_1|x_1)\delta_0(w_1)$ $p(u_1|x_1)$ and the transitions $p(v_{n+1},y_{n+1}|v_1^n,y_1^n)=p(v_{n+1}|v_n)$ $p(y_{n+1}|v_{n+1},y_1^n)$. The transitions $p(v_{n+1}|v_n)$ are given by $p(v_{n+1}|v_n)=p(x_{n+1}|v_n)p(w_{n+1}|x_{n+1},v_n)p(u_{n+1}|w_{n+1},x_{n+1},v_n)$. In the last equality one has $p(x_{n+1}|v_n)=p(x_{n+1}|x_n,u_n)$, $p(w_{n+1}|x_{n+1},v_n)=p(w_{n+1}|x_{n+1},x_n,w_n)$, and $p(u_{n+1}|w_{n+1},x_{n+1},v_n)=p(u_{n+1}|x_{n+1},u_n)$. Besides, $p(y_{n+1}|v_{n+1},y_1^n)=p(y_{n+1}|x_{n+1},y_{n-w_{n+1}+1}^n)$. Finally, one has

$$p(v_{n+1},y_{n+1}|v_1^n,y_1^n)=p(x_{n+1}|x_n,u_n)p(w_{n+1}|x_{n+1},x_n,w_n)\times p(u_{n+1}|x_{n+1},u_n)p(y_{n+1}|x_{n+1},y_{n-w_{n+1}+1}^n) \tag{4.1}$$

with the convention $p(y_{n+1}|x_{n+1},y_{n+1}^n)=p(y_{n+1}|x_{n+1})$.

As in (2.1) and (2.2), the four transitions in (4.1) can be written using the Kronecker functions. For example, if $u_n>0$, the probability for $x_{n+1}=x_n$ is one, which is written "$p(x_{n+1}|x_n,u_n)=\delta_{x_n}(x_{n+1})$ for $u_n>0$". One has

$$p(x_{n+1}|x_n,u_n)=\begin{cases}\delta_{x_n}(x_{n+1}) & \text{for } u_n>0\\ p(x_{n+1}|x_n) & \text{for } u_n=0\end{cases} \tag{4.2}$$

$$p(w_{n+1}|x_{n+1},x_n,w_n)=\begin{cases}\delta_{w_n+1}(w_{n+1}) & \text{for } w_n<S \text{ and } x_n=x_{n+1}\\ \delta_0(w_{n+1}) & \text{for } x_n\neq x_{n+1} \text{ or } w_n=S\end{cases} \tag{4.3}$$

$$p(u_{n+1}|x_{n+1},u_n)=\begin{cases}\delta_{u_n-1}(u_{n+1}) & \text{for } u_n>0\\ p(u_{n+1}|x_{n+1}) & \text{for } u_n=0\end{cases} \quad (4.4)$$

$$p(y_{n+1}|x_{n+1}=\omega_j,y_{n-w_{n+1}+1}^{n})=q^j(y_{n+1}|y_{n-w_{n+1}+1}^{n}) \quad (4.5)$$

However, let us underline the fact that no approximation is needed for $N\leq 1000$.

5. Parameter estimation with ICE

5.1. General ICE principle

The "Iterative Conditional Estimation" (ICE) method is based on the following principle [8,14,21]. Let $\theta=(\theta_1,\ldots,\theta_m)$ be the vector of all real parameters defining the distribution $p(v,y)$ of a couple of variables (V,Y), where Y is observed and V is hidden. Let $\hat{\theta}(v,y)$ be an estimator of θ defined from the complete data (v,y). ICE is an iterative method producing a sequence (θ^q) in the following way:

(i) initialize θ^0;
(ii) compute $\theta_i^{q+1}=E[\hat{\theta}_i(V,Y)|Y=y,\theta^q)$ for the components θ_i for which this computation is workable;
(iii) for other components θ_i, simulate $v^q,\ldots,v_l^q$ according to $p(v|y,\theta^q)$ and set $\theta_i^{q+1}=[\hat{\theta}_i(v_1^q,y)+\cdots+\hat{\theta}_i(v_l^q,y)]/l$.

One notices that ICE is applicable under two very slight hypotheses: existence of an estimator $\hat{\theta}(v,y)$ from the complete data, and the ability to simulate V according to $p(v|y)$. The first hypothesis is not really a constraint because if one is not able to estimate θ from complete data (v,y), there is no point in searching for an estimator from incomplete ones y. The second hypothesis is verified once the distribution $p(v|y)$ is a Markov chain distribution, which will be verified in this paper.

Remark 5.1. Let us underline the difference between ICE and the classical "Expectation-Maximization" (EM) method. The intuitive reason behind the ICE principle is the following. To simplify, let us suppose that $\theta\in R$, and let (X,Y) be a couple of random variables whose distribution depends on θ, and in which Y is observable and X is hidden. In general, one can estimate θ from complete data (X,Y) with an estimator $\hat{\theta}(X,Y)$, whose efficiency is often measured by the mean square error $E_\theta[(\theta-\hat{\theta}(X,Y))^2]$. As X is not available, the idea is to approximate $\hat{\theta}(X,Y)$ by some function of Y, and the best approximation, in the sense of mean square error, is the conditional expectation $(Y)=E_\theta[\hat{\theta}(X,Y)|Y]$. Thus, on the one hand, $\hat{\theta}(X,Y)$ is close to θ in the mean square error sense and, on the other hand, $\tilde{\theta}(Y)$ is close to $\hat{\theta}(X,Y)$ according to the same criterion. In other words, the possibly interesting "mean square error" properties of $\hat{\theta}(X,Y)$ are saved as far as possible by using $\tilde{\theta}$. Of course, $\tilde{\theta}$ is no longer an estimator because it does depend on θ, which leads to the ICE principle. The EM principle is $\theta^{q+1}=\operatorname{argmax}_\theta E[\log[p_\theta(X,Y)]|Y=y,\theta^q)$, and it has been successfully applied in different classical hidden semi-Markov models. Thus, ICE and EM follow different principles; however, they can produce the same sequence of parameters in some particular situations [7].

5.2. ICE in LDGN-NHSMC

Let us consider the general parameterized LDGN-NHSMC (X,U,Y) introduced in Section 4.2. To simplify, one considers the classical case where $p(y|x,u)=p(y|x)$, but the general case can be dealt with in exactly similar way. Thus, for K classes there are K LDGNs, each of which is defined by the mean $M^j=(m^j,\ldots,m^j)\in R^N$ and the variance-covariance matrix $\Gamma^j=[\gamma_{nr}^j]_{1\leq n,r\leq N}$, with $\gamma_{nr}^j=\gamma^j(n-r)=c^j(|n-r|+1)^{-\alpha^j}$. The parameters defining the distribution of the stationary Markov chain $V=(X,U)$ are those defining the distribution $p(v_1,v_2)=p(x_1,u_1,x_2,u_2)$. As the LDGN-NHSMC (V,Y) is a particular Markov chain hidden with LDGN, the ICE proposed in [21] is applicable in quite a straightforward way. Let us briefly remember how it works:

(i) initialization θ^0;
(ii) for θ^q, the next value $p^{q+1}(v_1,v_2)$ of the distribution $p(v_1,v_2)$ is given by the conditional expectation, which is computable, by the very classical empirical estimate. To re-estimate the parameters $(m^1, c^1, \alpha^1),\ldots,(m^K, c^K, \alpha^K)$ (remember that there are K classes), one samples $v^q=(x^q, u^q)$ according to $p(v|y, \theta^q)$ (just one value of v will be used, which means that in (iii) $l=1$). Then the data y are transformed into data y^* in such a way that for each $n=1,\ldots, N-1$ and $j=1,\ldots, K$ the distribution of $p(y_n^*,y_{n+1}^*|x_n=\omega_j,x_{n+1}=\omega_j)$ is Gaussian with the mean (m^j, m^j) and the variance–covariance matrix

$$\begin{bmatrix}c^j & c^j 2^{-\alpha^j}\\ c^j 2^{-\alpha^j} & c^j\end{bmatrix}$$

(the transformation used is based on the current parameter θ^q). Then the Gaussian parameters of $p(y_n^*,y_{n+1}^*|(x_n,x_{n+1})=(\omega_j,\omega_j))$ are estimated from (x^q,y^*) by the classical estimators, which gives the next parameters $(m^j)^{q+1}$, $(c^j)^{q+1}$, $(\alpha^j)^{q+1}$.

Such a method, which gave satisfying results in [21], can easily be extended to the case where the noise distribution $p(y|v)=p(y|x,u)$ depends on both (x,u): the NSMC (X,U) would be the same and there would be KL triplets (m^j, c^j, α^j) instead of K.

5.3. ICE in bounded memory model

Let us detail ICE in the bounded and stochastic memory model proposed in Section 4.3, which will be used in the experiments below. One adds the chain W to the model discussed in the previous sub-section, which does not modify the parameters. Thus one has to estimate the distribution $p(x_1, u_1, x_2, u_2)$ and $(m^1, c^1, \alpha^1), \ldots, (m^K, c^K, \alpha^K)$. The difference with the previous model is that searching for y^* is not necessary. In fact, $p(y_n,y_{n+1}|x_n,x_{n+1})$ is here Gaussian with the mean (m^j,m^j) and the variance–covariance matrix

$$\begin{bmatrix}c^j & c^j 2^{-\alpha^j}\\ c^j 2^{-\alpha^j} & c^j\end{bmatrix}$$

if $x_n=x_{n+1}=\omega_j$, and $p(y_n,y_{n+1}|x_n,x_{n+1})=p(y_n|x_n)p(y_{n+1}|x_{n+1})$ if $x_n\neq x_{n+1}$.

The distribution $p(x_1, u_1, x_2, u_2)$ is estimated from the complete data $T=(X, U, W, Y)$ with the classical empirical estimator

$$\hat{p}(x_1,u_1,x_2,u_2)=\frac{1}{N-1}\sum_{n=1}^{N-1} I(x_n=x_1,u_n=u_1,x_{n+1}=x_2,u_{n+1}=u_2) \tag{5.1}$$

where the function I is defined by $I(a=b)=1$ if $a=b$, and 0 otherwise.

Taking the conditional expectation of (5.1) gives

$$p^{(q+1)}(x_1,u_1,x_2,u_2)=\frac{1}{N-1}\sum_{n=1}^{N-1} p(x_n=x_1,u_n=u_1,x_{n+1}=x_2,u_{n+1}=u_2|y,\theta_q) \tag{5.2}$$

which is computable; in fact, $p(x_n,u_n,x_{n+1},u_{n+1}|y,\theta_q)=p(v_n,v_{n+1}|y,\theta_q)$ are computable with (i)–(v), Section 3.1.

The "noise parameters" $(m^1, c^1, \alpha^1),\ldots,(m^K, c^K, \alpha^K)$ will be estimated from $T=(X, U, W, Y)$ by the following classical estimators. For $j=1,\ldots, K$, let $N_j=\sum_{n=1}^{N} I(x_n=\omega_j)$ and $N_{jj}=\sum_{n=1}^{N-1} I(x_n=\omega_j, x_{n+1}=\omega_j)$. One sets $\hat{m}^j=(1/N_j)\sum_{n=1}^{N} y_n I(x_n=\omega_j)$, $\hat{c}^j=(1/N_j)\sum_{n=1}^{N}(y_n-\hat{m}^j)^2 I(x_n=\omega_j)$ and $\hat{\alpha}^j=-\log(\hat{\gamma}^j)/\log(2)$, where $\hat{\gamma}^j=(1/N_{jj})\sum_{n=1}^{N-1}(y_n-\hat{m}_1^j)(y_n-\hat{m}_2^j)I(x_n=\omega_j,x_{n+1}=\omega_j)$, with $\hat{m}_1^j=(1/N_{jj})\sum_{n=1}^{N-1} y_n I(x_n=x,x_{n+1}=x)$ and $\hat{m}_2^j=(1/N_{jj})\sum_{n=1}^{N-1} y_{n+1}\ I(x_n=x,x_{n+1}=x)$.

Concerning the noise parameters, the conditional expectation is computable for none of them, and one has to use the sampled $v^q=(x^q, u^q)$. In experiments below the initialization is obtained from the segmentation by the classical k-means method.

6. Experiments

The new "hidden semi-Markov chains with long dependence" noise (LDGN-NHSMC) model extends, on the one hand, the classical "hidden semi-Markov chains" (HSMC) and, on the other hand, the "Markov chains hidden with long dependence" noise (LDGN-HMC). The aim of this section is to test the interest of these two generalizations in an unsupervised data segmentation framework. The four following models will be compared: the very classical "hidden Markov chain" (HMC) given by (1.1), the "Markov chain hidden with Gaussian long dependence noise" LDGN-HMC, the "new hidden semi-Markov chain" NHSMC, and the new LDGN-NHSMC proposed. Numerous tests have been performed and some of them, which are representative of different other experiments, are presented below.

Four series of experiments are proposed:

In the first series, the data are produced by NHSMC, and the question is to study whether using the new more complex LDGN-NHSMC degrades the results or not.

The second series is devoted to the converse problem: when data are sampled according to LDGN-NHSMC, how do NHSMC and LDGN-HMC work?

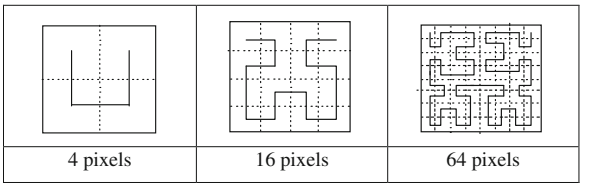

Fig. 1. Construction of a Hilbert–Peano scan used to represent realizations of mono-dimensional sequences as bi-dimensional images

In the third series one uses data produced by none of the four models.

In the three first series the noise is sampled and thus its level is controlled.

Finally, in the fourth series one considers a real radar image on which one applies HMC, LDGN-HMC, and LDGN-NHSMC.

To illustrate the results one will use images of size $N=128\times128$. Such a bi-dimensional set of pixels is transformed into a mono-dimensional set using a Hilbert–Peano scan [14], presented in Fig. 1, which gives a mono-dimensional chain. Such a representation is quite pleasant because it allows one to appreciate visually the degree of the noise, and the quality difference between two segmentation results as well. However, let us insist on the fact that in the three following sub-sections this is only a representation, and the problem dealt with is the problem of mono-dimensional chains. Therefore, the results presented are of interest in any area mentioned in the Introduction. However, it can also be used in image segmentation, which is dealt with in Section 6.4.

In all experiments presented in this section one hundred iterations were used for ICE.

6.1. HSMC data segmented with LDGN-NHSMC

Let (X, U, Y) be an NHSMC, with $K=2$ and $L=10$. The means of the Gaussian distributions $p(y_n|x_n=\omega_1)$ and $p(y_n|x_n=\omega_2)$ are equal to 1 and 2, respectively, and their common variance is equal to 20. In the following, one will adopt the notations $p(y_n|x_n=\omega_1)\sim N(1,20)$ and $p(y_n|x_n=\omega_2)\sim N(2,20)$. The distribution of $p(u_{n+1}|x_{n+1},u_n=0)$ is uniform on Λ for each x_{n+1}, and $p(x_n,x_{n+1}|u_n=0)=0.4995$ for $x_n=x_{n+1}$, and $p(x_n,x_{n+1}|u_n=0)=0.0005$ for $x_n\neq x_{n+1}$. The obtained realization $Y=y$, presented in Fig. 2, is then segmented by three methods. The first one is the MPM method based on the true NHSMC model and the true parameters; thus the result is used as the reference. The second method is the MPM unsupervised method based on the classical HSMC and ICE, while the third method is the MPM unsupervised method based on the new LDGN-NHSMC model, with $S=50$, and the related ICE. The aims of this experiment are, on the one hand, to show the robustness of the LDGN-NHSMC model and on the other hand, to see how the new model manages the independent noise.

According to the results presented in Fig. 2, the new model gives comparable results to those obtained with

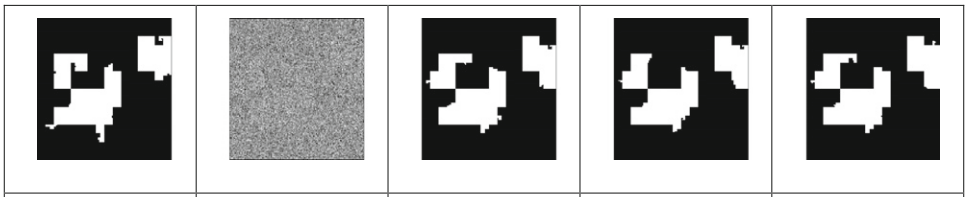

Fig. 2. Segmentation of NHSMC data according to three methods.

Table 1
Estimation of the parameters of NHSMC and LDGN-HSMC from data produced by NHSMC. τ: error rate of wrongly classified pixels.

	Real parameters		NHSMC		LDGN-HSMC	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	1.00	2.00	1.01	2.04	0.98	1.97
σ^2	20.00	20.00	19.81	20.71	19.84	20.46
α	–	–	–	–	15.28	5.95
τ (%)	3.74		4.55		4.57	

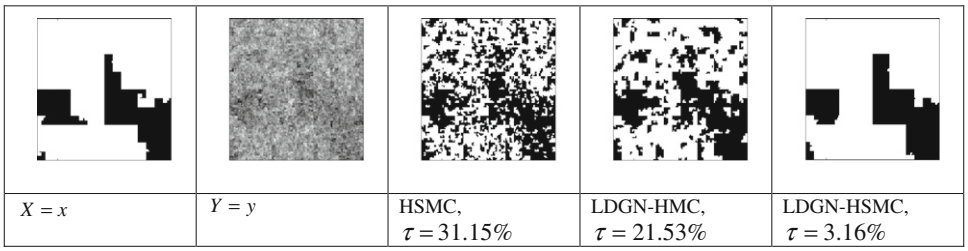

Fig. 3. Three unsupervised segmentations, based on HSMC, LDGN-HMC, and LDGN-HSMC, of data simulated according to HSMC-LDN. τ: error rate of wrongly classified pixels.

NHSMC. This is probably due to the good behaviour of the noise parameters estimation (see estimates in Table 1). Otherwise, one can see that the level of the noise is really high. In fact, when it is lower the three methods give almost identical results and thus one had to increase it to obtain some difference. This means, on the one hand, that ICE works very well and, on the other hand, that using LDGN-NHSMC instead of NHSMC hardly degrades the results. In other words, when the data do follow NHSMC, the more general LDGN-NHSMC can be used in an unsupervised manner without any risk.

6.2. LDGN-NHSMC data segmented with NHSMC and LDGN-HMC

Here one considers the converse problem. When data follow the new LDGN-NHSMC model, can either NHSMC or LDGN-HMC provide results comparable to those obtained with the LDGN-NHSMC model? In other words, when data have both "long dependence noise" and "semi-Markovianity of the hidden chain" properties, can assuming just one of them be sufficient in unsupervised segmentation? Let us consider the same semi-Markov chain (X, U) as above. For the long dependence noise, the means are, respectively, equal to 1 and 2, the common variance is equal to 1, and $\alpha^1=\alpha^2=\alpha=0.5$.

Table 2
Estimation of the noise parameters in the three models NHSMC, LDGN-HMC, and LDGN-NHSMC.

	Real parameters		NHSMC		LDGN-HMC		LDGN-NHSMC	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	1.00	2.00	0.97	2.44	1.08	2.22	1.03	1.98
σ^2	1.00	1.00	0.59	0.56	0.83	0.78	0.96	0.93
α	0.50	0.50	–	–	0.69	0.72	0.62	0.61

According to the results presented in Fig. 3, neither NHSMC nor LDGN-HMC can compete with LDGN-NHSMC when data are produced by the latter. The difference in error ratios is very large, which means that LDGN-NHSMC is a really significant extension of both NHSMC and LDGN-HMC Fig. 3. Besides, in spite of the high noise level one can notice that ICE works quite well (Table 2).

6.3. Hand written data corrupted with spatially correlated noise

The aim of this series is to consider data provided by none of the studied models. Let us consider a hand-written image "target" $X=x$ of size 128×128 presented in Fig. 4. One will consider two correlated noises obtained by the spatially "mobile mean": first, for each pixel s of the image, one samples an independent Gaussian noise $B_s \sim N(0,1)$. Then one computes $C_s = a\sum_{t \in V_s} B_t$, where V_s is a window of size 25×21 centered at s, and a is chosen such that $Var[C_s]=1$.

Let us consider two examples. In the first one one sets $Y_s=C_s+1$ if $X_s=\omega_1$, and $Y_s=C_s+2$ if $X_s=\omega_2$. Thus one obtains a correlated Gaussian distribution $p(y|x)$, with $p(y_s|x_s=\omega_1) \sim N(1,1)$ and $p(y_s|x_s=\omega_2) \sim N(2,1)$. As above, one considers $L=10$ for the sequence U and $S=50$ for the sequence W.

Bearing in mind that the mono dimensional chain Y_1, …, $Y_{128 \times 128}$ is obtained from the bi-dimensional set of pixels by the Hilbert–Peano scan presented in Fig. 1, one can say that the structure of the correlations of the Gaussian distribution $p(y|x)$ seen as the distribution of a mono-dimensional chain is very complex and undoubtedly quite different from any "long dependence" form previously considered. Moreover, given the form of the Hilbert–Peano scan the chain (X,Y) cannot be considered to be stationary. Thus one sees that in this sub-section the "Hilbert–Peano" representation is used to simply produce data which suit none of the considered models.

The results of different unsupervised segmentation methods are presented in Fig. 4, while the estimates of the noise parameters obtained with ICE are given in Table 3. Concerning the initialization of ICE obtained with the classical "C-means" classification algorithm, one finds $m_1^0=0.75$, $m_2^0=2.66$, $(\sigma_1^2)^0=0.48$, and $(\sigma_1^2)^0=0.47$. Thus the ICE estimation of the means is excellent in both models LDGN-HMC, LDGN-NHSMC, while the estimation of the variances is less efficient.

It appears that HMC cannot manage the noise correlation and gives quite poor results. Introducing long dependence noise in the hidden Markov chains improves things and the error rate passes from 30.0% to 22.8%. However, the really interesting result is that introducing LDGN-NHSMC still significantly improves the results obtained with LDGN-HMC, with the error rate passing from 22.8% to 14.2%. This clearly shows that using both aspects "long dependence noise" and "semi-Markovianity" simultaneously can be of great interest. This also shows that these two aspects model different things and that one cannot replace the other.

In the second example, one sets $Y_s=C_s$ if $X_s=\omega_1$, and $Y_s=\sqrt{5}C_s$ if $X_s=\omega_2$. Thus one obtains a correlated Gaussian distribution of $p(y|x)$, with all means null and the variance of $p(y_s|x_s=\omega_1)$ equal to 1, while the variance of $p(y_s|x_s=\omega_2)$ is equal to 5. As above, $Y=y$ is segmented by using the three methods HMC, LDGN-HMC and LDGN-NHSMC. As above, one considers $L=10$ and $S=50$.

The results are presented in Fig. 5. As above, when the noise is not strong enough the results of the three methods considered are comparable, and thus one had to consider a rather strong noise, as shown in Fig. 5. One can formulate the same conclusions as in the first example: the use of LDGN-NHSMC can greatly improve the results obtained with LDGN-HMC and HMC. Concerning the parameter estimation, one can notice that ICE encounters some problems in correctly estimating of the variances, while means are well estimated and the estimation of α gives comparable results in both LDGN-HMC and LDGN-NHSMC cases (Table 4).

Concerning the computation time, the classical HMC based method takes about 10 s, LDGN-HMC takes about 25 min, and LDGN-NHSMC takes about 45 min.

6.4. Real radar image segmentation

Finally, let us consider a real radar SAR airborne image. This image, seen as a realization $Y=y$, is presented in Fig. 6. The size here is 256×256 pixels. The set of pixels is converted into mono-dimensional sequence using the Hilbert–Peano scan and the observed sequence is then segmented by three methods based on HMC, LDGN-HMC and LDGN-NHSMC, respectively. The results obtained are presented in Fig. 6, and the parameter's estimates are given in Table 5.

Table 3
Estimation of the noise parameters in HMC, LDGN-HMC, and LDGN-NHSMC.

	HMC		LDGN-HMC		LDGN-NHSMC	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	0.75	2.27	1.02	2.07	0.99	2.03
σ^2	0.52	0.52	0.60	0.65	0.61	0.66
α	–	–	1.38	1.31	1.43	1.25

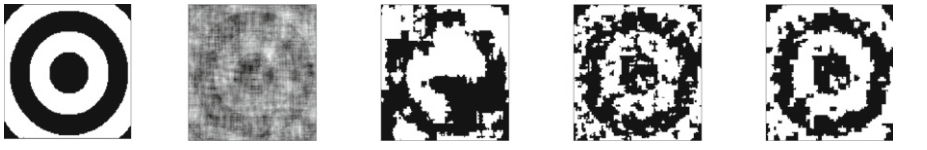

Fig. 4. Unsupervised segmentations of a hand written image noisy with spatially correlated noise using HMC, LDGN-HMC, and LDGN-NHSMC. τ: error rate of wrongly classified pixels.

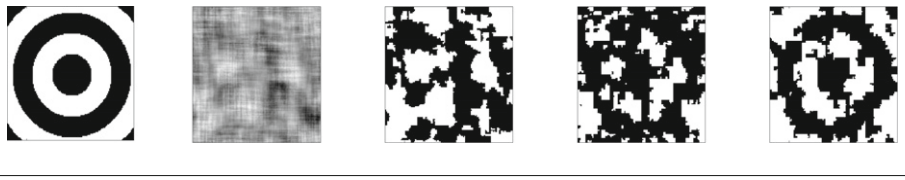

Fig. 5. Three unsupervised segmentations of a hand written image noisy with spatially correlated noise.

Table 4
Estimation of the parameters in the three models HMC, LDGN-HMC, and LDGN-NHSMC.

	HMC		LDGN-HMC		LDGN-NHSMC	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	1.29	−1.09	0.07	−0.03	−0.04	0.10
σ^2	1.02	1.30	1.38	4.03	1.74	3.56
α	–	–	0.15	0.13	0.12	0.14

Table 5
Estimation of the parameters in the three models HMC, LDGN-HMC, and LDGN-NHSMC in the real SAR image presented in Fig. 6.

	HMC			LDGN-HMC			LDGN-NHSMC		
	ω_1	ω_2	ω_3	ω_1	ω_2	ω_3	ω_1	ω_2	ω_3
m	9.51	20.12	4.25	12.78	20.63	4.77	17.65	20.57	6.25
σ^2	18.37	5.46	8.74	31.89	4.35	11.56	7.43	4.40	14.28
α	–	–	–	0.13	2.01	5.86	0.99	1.69	1.84

In LDGN-HMC case one has taken S=10, and one has taken S=L=10 in the LDGN-NHSMC case. One assumes the existence of three classes: ω_1 (green) is the river, ω_2 (blue) is the vegetation, and ω_3 (brown) is what remains in the image.

One can notice that the LDGN-NHSMC based segmentation method clearly provides the best results. In particular, it better finds the river and does not confuse the river with the vegetation, as HMC and LDGN-HMC based methods do.

Remark 6.1. The hidden Markov fields (HMF) are intuitively better suited to dealing with unsupervised image segmentation and let us underline the fact that the new LDGN-NHSMC model is not proposed to compete with HMF in image segmentation problems. As LDGN-NHSMCs extend LDGN-HMCs, NHSMCs, and HMCs, they are automatically of interest in all contexts the latter three models are, provided however the parameters are estimated efficiently. The main purpose of the Sections 6.1 and 6.2 was to show that ICE works well, even under very high noise. However, the results presented in the Section 6.3 and the present one present a double interest.

First, given the form of the Hilbert–Peano scan the random chain relating to the mono-dimensional sequence so obtained is very irregular and its distribution is probably very far from the distributions of all models studied. In addition, it is well known that SAR images are difficult to segment. This shows the very good robustness of the new LDGN-NHSMC model, at least with respect to the classical HMC and LDGN-HMC ones.

Second, even in unsupervised image segmentation context the use of HMC can be of interest with respect to the use of HMF in some situations [14]. In particular, using HMF is very time consuming and thus HMC can be used to initialize the HMF-based processing. Thus LDGN-NHSMC can be used instead of HMC in such situations, probably with better efficiency.

7. Conclusions

This paper contains two novelties. First, a new family of semi-Markov chains has been proposed. In the case in which they are hidden with the classical independent noise, using the models from this family is as easy as using HMC, both in terms of computational cost and the parameter estimation. Second, the classical hidden model relating to this family has been extended to the long dependence noise case. A parameter estimation method has been defined and different experiment results have been provided.

As a general conclusion one can say, according to different experiments results, that the new semi-Markov chain hidden with the long dependence noise model proposed, turns out to be of interest, when unsupervised segmentation is concerned, with respect to classical simpler models. More precisely, it is possible to put forth the following points:

(i) The main result is that there are situations in which the data are very noisy, they suit none of the considered models, and in which the new model clearly gives the best results. In other words, both the "long dependence" of the noise and the "semi-Markovianity" of the hidden chain are of importance and can strongly contribute to the improvement of results obtained with simpler classical models.
(ii) The parameter estimation method proposed, which is of the "Iterative Conditional Estimation" (ICE) kind, is well suited to the problem, even in very highly noisy cases.
(iii) The proposed models and related unsupervised processing is immediately applicable to different areas mentioned in the Introduction. In particular, the last example of the previous section shows that its application in image segmentation is promising.

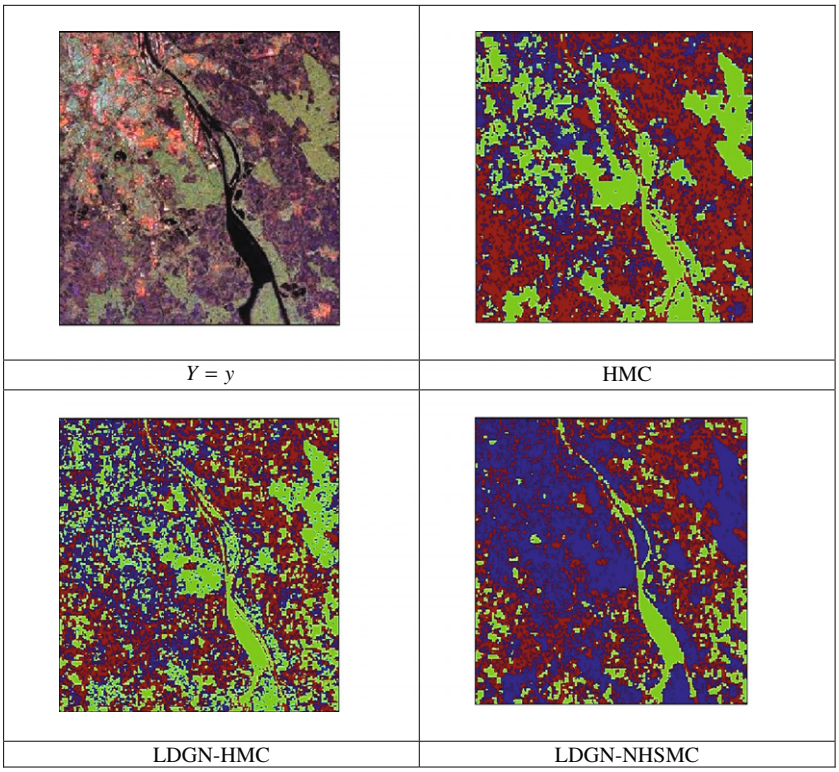

Fig. 6. Three unsupervised segmentations of a real SAR airborne image of the "Rhone Valley", France. ω_1 (green) is the river, ω_2 (blue) is the vegetation, and ω_3 (brown) is what remains. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Among perspectives, let us mention the possible use of copulas, to generalize the Gaussian noise used in this paper to noises of any marginal distributions [3]. Another perspective could be applying LDGN-NHSMC to 3D images [2].

References

[1] V.S. Barbu, N. Limnios, Semi-Markov chains and hidden semi-Markov models toward applications, Lecture Notes in Statistics, Springer, 2008.
[2] S. Bricq, C. Collet, J.-P. Armspach, Unifying framework for multimodal brain MRI segmentation based on hidden Markov chains, Medical Image Analysis 16 (6) (2008) 639–652.
[3] N. Brunel, W. Pieczynski, Unsupervised signal restoration using hidden Markov chains with copulas, Signal Processing 85 (12) (2005) 2304–2315.
[4] J. Bulla, I. Bulla, Stylized facts of financial time series and hidden semi-Markov models, Computational Statistics and Data analysis 51 (4) (2006) 2192–2209.
[5] O. Cappé, E. Moulines, T. Ryden, Inference in Hidden Markov Models, Springer Series in Statistics, Springer, 2005.
[6] M. Coulon, M. Chabert, A. Swami, Detection of multiple changes in fractional integrated ARMA Processes, IEEE Transactions on Signal Processing 57 (1) (2009) 48–61.
[7] J.-P. Delmas, An equivalence of the EM and ICE algorithm for exponential family, IEEE Transactions on Signal Processing 45 (10) (1997) 2613–2615.
[8] S. Derrode, W. Pieczynski, Signal and image segmentation using pairwise Markov chains, IEEE Transactions on Signal Processing 52 (9) (2004) 2477–2489.
[9] F.X. Diebold, A. Inoue, Long memory and regime switching, Journal of Econometrics 105 (1) (2001) 131–159.
[10] P. Doukhan, G. Oppenheim, M.S. Taqqu (Eds.), Theory and Applications of Long-Range Dependence, Birkhäuser, 2003.
[11] Y. Ephraim, Hidden Markov processes, IEEE Transactions on Information Theory 48 (6) (2002) 1518–1569.
[12] S. Faisan, L. Thoraval, J.-P. Armspach, M.-N. Metz-Lutz, F. Heitz, Unsupervised learning and mapping of active brain functional MRI signals based on hidden semi-Markov event sequence models, IEEE Transactions on Medical Imaging 24 (2) (2005) 263–276.
[13] J.D. Fergusson, Variable duration models for speech, in: Symposium Application of Hidden Markov Models to Text and Speech, Institute for Defense Analyses, Princeton, NJ, 1980, pp. 143–179.
[14] R. Fjørtoft, Y. Delignon, W. Pieczynski, M. Sigelle, F. Tupin, Unsupervised segmentation of radar images using hidden Markov chains and hidden Markov random fields, IEEE Transactions on Geoscience and Remote Sensing 41 (3) (2003) 675–686.
[15] S. Forchhammer, J. Rissanen, Partially hidden Markov models, IEEE Transactions on Information Theory 42 (4) (1996) 1253–1256.
[16] M. Gales, S. Young, The application of hidden Markov models in speech recognition, Now Publishers, World Scientific, 2008.
[17] Y. Guédon, Estimating hidden semi-Markov chains from discrete sequences, Journal of Computational and Graphical Statistics 12 (3) (2003) 604–639.
[18] J. Janssen, R. Manca, Applied semi-Markov processes, Springer, New York, 2006.
[19] T. Koski, Hidden Markov models for bioinformatics, Kluwer Academic Publishers, Netherlands, 2001.
[20] A. Kundu, Y. He, M.-Y. Chen, Alternatives to variable duration HMM in handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1275–1280.
[21] P. Lanchantin, J. Lapuyade-Lahorgue, W. Pieczynski, Unsupervised segmentation of triplet Markov chains with long-memory noise, Signal Processing 88 (5) (2008) 1134–1151.
[22] J. Lapuyade-Lahorgue, W. Pieczynski, Partially Markov models and unsupervised segmentation of semi-Markov chains hidden with long dependence noise, in: International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2007), Chania, Greece, May 2007.

[23] J. Lapuyade-Lahorgue, Sur diverses extensions des chaînes de Markov cachées avec application au traitement des signaux radar, Ph.D. Dissertation, December 2008.
[24] R.S. Mamon, R.J. Elliott, Hidden Markov Models in Finance, Springer, 2007.
[25] M.D. Moore, M.I. Savic, Speech reconstruction using a generalized HSMM (GHSMM), Digital Signal Processing 14 (1) (2004) 37–53.
[26] G. Nuel et, B. Prum, Analyse statistique des séquences biologiques: Modélisation markovienne, alignements et motifs, Hermes, Paris, 2007.
[27] M. Ostendorf, V. Dikalakis, O. Kimball, From HMMs to segment models, IEEE Transactions on Speech and Audio Processing 4 (5) (1996) 360–378.
[28] W. Pieczynski, Pairwise Markov chains, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5) (2003) 634–639.
[29] W. Pieczynski, Multisensor triplet Markov chains and theory of evidence, International Journal of Approximate Reasoning 45 (1) (2007) 1–16.
[30] W. Pieczynski, F. Desbouvries, On triplet Markov chains, in: International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France, May 2005.
[31] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.
[32] M.J. Russel, A segmental HMM for speech pattern modeling, vol 2, ICASSP-93, 1993, pp. 499–502.
[33] M. S. Taqqu, Fractional Brownian motion and long-range dependence, in: Theory and Applications of Long-Range Dependence, Birhhäuser, 2003.
[34] X. Tan, H. Xi, Hidden semi-Markov model for anomaly detection, Applied Mathematics and Computation 205 (2) (2008) 562–567.
[35] S.V. Vaseghi, State duration modeling in hidden Markov models, Signal Processing 41 (1) (1995) 31–41.
[36] P. Wang, M. Theobald, Regime-switching volatility of six East Asian emerging markets, Research in International Business and Finance 22 (3) (2008) 267–283.
[37] L.H. Wang, Change-point detection with rank statistics in long-memory time-series models, Australian & New Zealand Journal of Statistics 50 (3) (2008) 241–256.
[38] B.-J. Yoon, P.P. Vaidyanathan, Context-sensitive hidden Markov models for modeling long-range dependencies in symbol sequences, IEEE Transactions on Signal Processing 54 (11) (2006) 4169–4184.
[39] S.-Z. Yu, H. Kobayashi, Practical implementation of an efficient forward–backward algorithm for an explicit-duration hidden Markov model, IEEE Transactions on Signal Processing 54 (5) (2006) 1947–1951.
[40] S.-Z. Yu, Hidden semi-Markov models, Artificial Intelligence 174 (2) (2010) 215–243.