Pairwise Markov Chains

Wojciech Pieczynski

Abstract—We propose a new model called a Pairwise Markov Chain (PMC), which generalizes the classical Hidden Markov Chain (HMC) model. The generalization, which allows one to model more complex situations, in particular implies that in PMC the hidden process is not necessarily a Markov process. However, PMC allows one to use the classical Bayesian restoration methods like Maximum A Posteriori (MAP), or Maximal Posterior Mode (MPM). So, akin to HMC, PMC allows one to restore hidden stochastic processes, with numerous applications to signal and image processing, such as speech recognition, image segmentation, and symbol detection or classification, among others. Furthermore, we propose an original method of parameter estimation, which generalizes the classical Iterative Conditional Estimation (ICE) valid for of classical hidden Markov chain model, and whose extension to possibly non-Gaussian and correlated noise is briefly treated. Some preliminary experiments validate the interest of the new model.

Index Terms—Bayesian restoration, hidden data, image segmentation, iterative conditional estimation, hidden Markov chain, pairwise Markov chain, unsupervised classification.

1 Introduction

The field of applications of Hidden Markov Models is extremely vast. Among this family of models, Hidden Markov Chains (HMC) are among the most frequently used. In the pattern recognition and image processing area, HMC can be used in image segmentation [5], [22], handwritten word recognition [4], [12], document image analysis, tumor classification, vehicle detection [1], acoustic musical signal recognition [24], or even gesture recognition [26]. Multisensor images can still be segmented using hidden Markov chains [15]. A priori, Hidden Markov Random Fields (HMRF) are better suited to deal with the image segmentation problem [3], [14], [18], although HMC-based segmentation methods can be competitive in some particular situations [25], and they are much faster than the HMRF-based ones. Other potential applications include speech recognition [8], [23], communications [16], or genome structure recognition [6]. We may also mention the quad tree model [17], which can be seen as more sophisticated hidden Markov chains, with applications to statistical image segmentation.

The success of such models is due to the fact that when the unobservable, or hidden, signal can be modeled by a finite Markov chain and when the noise is not too complex, then the signal can be recovered using different Bayesian classification techniques like Maximum A Posteriori (MAP), or Maximal Posterior Mode (MPM) [2], [13]. These restoration methods use the distribution of the hidden process conditional to the observations, which is called its "posterior" distribution.

Furthermore, such restoration techniques can be rendered unsupervised by applying some parameter estimation method, like Expectation-Maximization (EM) [2], [11], or Iterative Conditional Estimation (ICE) [9], [10], [15]. EM and ICE present some common properties [10], and ICE can be an alternative to EM in some complex situations occurring in image processing, like using hierarchical random models [19]. Let us mention that in images different noise contributions are not necessarily Gaussian and, thus, EM and ICE have been extended to "generalized" estimation methods, in which the very nature of different noise processes is searched in an unsupervised manner [9], [15], [20].

For instance, let us consider the following hidden Markov model: $X = (X_1, \ldots, X_n)$ is a Markov chain, with all X_i taking their values in the set of classes $\Omega = \{\omega_1, \ldots, \omega_k\}$, and $Y = (Y_1, \ldots, Y_n)$ is the process of observations, each Y_i taking their values in R. Thus, P_X is a Markov distribution and one has to define the distributions $P_Y^{X=x}$ of Y conditional to X in such a way that the posterior distribution $P_X^{Y=y}$ is still a Markov distribution. We shall insist that the Markovianity of $P_X^{Y=y}$ is essential to the successful application of Bayesian MAP or MPM restoration methods. Indeed, $P_X^{Y=y}$ is Markovian for a large family of distributions and this very fact is the origin of the success of the hidden Markov models. However, there also exist simple distributions $P_Y^{X=x}$ for which P_X is a Markov distribution, while $P_X^{Y=y}$ is no longer a Markovian one. For example, when $P_Y^{X=x}$ is Gaussian, the hypothesis $P_{Y_i}^{X=x} = P_{Y_i}^{X_i=x_i}$, which can be easily questioned, is frequently used to retain the Markovianity of $P_X^{Y=y}$.

The aim of this paper is to propose a model which is more general than the hidden Markov chain model and in which the posterior distribution $P_X^{Y=y}$ will always be a Markov chain distribution. The main objective is to allow one to consider more complex $P_Y^{X=x}$, which could possibly be better suited to different real data. As we will see in the following, the hypothesis $P_{Y_i}^{X=x} = P_{Y_i}^{X_i=x_i}$ above can, in particular, be relaxed. The idea is to directly consider the Markovianity of the couple (X, Y): such a model will be called "Pairwise Markov Chain" (PMC). The difference with the HMC is that the distribution P_X is not necessarily a Markov distribution, but $P_X^{Y=y}$ always is. This latter property allows one to use Bayesian restoration methods, like MPM or MAP.

Concerning the PMC parameter estimation problem, we show how ICE can be used in the Gaussian case and briefly indicate how it can be extended to the case in which the very form of the noise is not known and has to be searched, the latter being inspired from [20].

The paper is organized as follows: The PMC model is introduced in the next section and some basic properties are presented. Differences in practical calculus with respect to the HMC are specified in Section 3 and Section 4 is devoted to the parameter estimation problem. Some numerical results are presented in Section 5 and conclusions and perspectives are in Section 6.

2 Pairwise Markov Chains

Let us consider two sequences of random variables $X = (X_1, \ldots, X_n)$, and $Y = (Y_1, \ldots, Y_n)$. Each X_i takes its values in a set X and each Y_i takes its values in a set Y. Then, let $Z_i = (X_i, Y_i)$ be the "pairwise" variable at the point i, and let $Z = (Z_1, \ldots, Z_n)$ be the "pairwise" process corresponding to two processes X and Y. We will assume that different probability distributions corresponding to the different variables have densities with respect to some measures. For simplicity, we will denote these different densities by a same letter p. For instance, $p(x)$, $p(x_i)$, $p(x_i, x_{i+1})$, $p(z_i) = p(x_i, y_i)$ will be the densities of the distributions of X, X_i, (X_i, X_{i+1}), and $Z_i = (X_i, Y_i)$, respectively. The conditional densities will still be denoted by p: $p(x_{i+1}|x_i)$ will be the density of the distribution of X_{i+1} conditional on $X_i = x_i$, $p(y|x)$ will be the density of the distribution of Y conditional to $X = x$, etc. We do not specify the measures for the different densities because it is not necessary for what follows and, thus, this lack of specification provides a certain generality of the framework. Some of classical measures will be specified in the examples.

Definition 2.1. *Z will be called a Pairwise Markov Chain (PMC) associated with X and Y if its distribution may be expressed as*

$$p(z) = \frac{p(z_1, z_2)p(z_2, z_3)\ldots p(z_{n-1}, z_n)}{p(z_2)p(z_3)\ldots p(z_{n-1})}, \tag{2.1}$$

• *The author is with the Institut National des Télécommunications, Département CITI 9, rue Charles Fourier, 91000 Evry, France. E-mail: Wojciech.Pieczynski@int-evry.fr.*

Manuscript received 19 Dec. 2000; revised 31 July 2001; accepted 12 June 2002.
Recommended for acceptance by A. Kundu.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 113324.

where $p(.)$ are probability densities with respect to some measures. Furthermore, for $2 \leq i \leq n-1$, $p(z_i)$ is the marginal distribution of $p(z_i, z_{i+1})$ and it also is the marginal distribution of $p(z_{i-1}, z_i)$. Of course, Z is then a Markov chain, but we will keep the definition above because of its usefulness in the following.

Thus, the distribution of a pairwise Markov chain is given by the densities $p(z_1, z_2), \ldots, p(z_{n-1}, z_n)$. The PMC will be called "stationary" when these $n-1$ densities are equal. The distribution of a stationary PMC is thus given by a density on $\mathrm{Z}^2 = \mathrm{X}^2 \times \mathrm{Y}^2$ with respect to some measure.

The following proposition specifies some useful properties of PMC.

Proposition 2.1. *Let Z be a Pairwise Markov Chain (PMC) associated with X and Y. We have the following:*

1. *$p(y|x)$ and $p(x|y)$ are Markov chains;*
2. *the distribution of (Z_i, Z_{i+1}) is given by the density $p(z_i, z_{i+1})$.*

Proof.

1. Z being a Markov chain, we have $p(z) = p(z_1)p(z_2|z_1) \ldots p(z_n|z_{n-1})$ and, so,

$$p(y|x) = \frac{p(z)}{p(x)} = \frac{p(x_1, y_1)p(x_2, y_2|x_1, y_1) \ldots p(x_n, y_n|x_{n-1}, y_{n-1})}{p(x)}.$$

   As x is constant in the equalities above, we recognize a Markovian form of the density $p(y_1, \ldots, y_n|x)$. In the same way, $p(x|y)$ is Markovian.
2. Upon integrating (2.1) with respect to $z_1, \ldots, z_{m-1}$ on the one hand, and with respect to $z_n, \ldots, z_{l+1}$ on the other hand, we notice that $p(z_m, \ldots, z_l)$, which is a marginal distribution of $p(z)$ defined by (2.1), retains the structure of a PMC:

$$p(z_m, \ldots, z_l) = \frac{p(z_m, z_{m+1}) \ldots p(z_{l-1}, z_l)}{p(z_{m+1}) \ldots p(z_{l-1})}. \tag{2.2}$$

   Point 2 of the proposition follows by setting $l = m+1$.□

Example 2.1. Let us consider a classical hidden Markov chain with independent noise: $\mathrm{X} = \{\omega_1, \ldots, \omega_k\}$ is a finite set of classes, X a classical Markov chain on X, and $\mathrm{Y} = R$ is the set of real numbers. Furthermore, the random variables (Y_i) are independent conditionally to X and the distribution of each (Y_i) conditional to X is equal to its distribution conditional to X_i, given by $p(y_i|x_i)$. Assume that $p(y_i|x_i)$ is Gaussian density. The distribution of Z is then given by

$$p(z) = p(x, y) = p(x_1)p(y_1|x_1)p(x_2|x_1)p(y_2|x_2) \ldots p(x_n|x_{n-1})p(y_n|x_n).$$

Otherwise, we have $p(x_{i+1}|x_i) = \frac{p(x_i, x_{i+1})}{p(x_i)}$ and, thus, we have a pairwise chain defined by

$$p(z_i, z_{i+1}) = p(x_i, x_{i+1})p(y_i|x_i)p(y_{i+1}|x_{i+1}).$$

As specified above, p designates different densities with respect to some measures on different subsets of $\mathrm{X}^n \times \mathrm{Y}^n$. Here, we have two different measures: a counting measure on X^n and the Lebesgue measure on Y^n. So, when the vector x, or some of its subvectors, are concerned different p are simply probabilities, and, when the vector y, or some of its subvectors, are concerned, different p are densities with respect to the Lebesgue measure. When both are concerned, as in (2.1), p is a density with respect to a product of a counting measure with a Lebesgue measure.

In the following two propositions, we identify some conditions on $p(z_1, z_2)$ under which the hidden process X is a Markov chain.

Proposition 2.2. *Let Z be a stationary PMC associated with X and Y and defined by $p(z_1, z_2)$. The hypothesis (H) below*

$$p(y_1|x_1, x_2) = p(y_1|x_1) \tag{H}$$

implies that X is a Markov chain. Furthermore, the distribution of the Markov chain X is

$$p(x) = \frac{p(x_1, x_2) \ldots p(x_{n-1} \ldots x_n)}{p(x_2) \ldots p(x_{n-1})}.$$

Of course, as the model is symmetric with respect to x and y, an analogous result remains true upon exchanging x and y.

Proof. Putting $x^n = (x_1, \ldots, x_n)$, $y^n = (y_1, \ldots, y_n)$, and $z^n = (z_1, \ldots, z_n)$, we have

$$\begin{aligned} p(x^n, y^n) = p(z^n) &= \frac{p(z_1, z_2) \ldots p(z_{n-1}, z_n)}{p(z_2) \ldots p(z_{n-1})} \\ &= \frac{p(x_1, x_2)p(y_1, y_2|x_1, x_2) \ldots p(x_{n-1}, x_n)p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(x_2)p(y_2|x_2) \ldots p(x_{n-1})p(y_{n-1}|x_{n-1})} \\ &= \left[\frac{p(x_1, x_2) \ldots p(x_{n-1}, x_n)}{p(x_2) \ldots p(x_{n-1})}\right] \\ &\quad \left[\frac{p(y_1, y_2|x_1, x_2) \ldots p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(y_2|x_2) \ldots p(y_{n-1}|x_{n-1})}\right] \\ &= q(x^n)\varphi_n(x^n, y^n). \end{aligned}$$

Thus, $p(x^n) = \int_{\mathrm{Y}^n} p(x^n, y^n) d\nu^n(y^n) = q(x^n)$ will be implied by $\int_{\mathrm{Y}^n} \varphi_n(x^n, y^n) d\nu^n(y^n) = 1$ and, so, it remains to show the latter. We have

$$\begin{aligned} &\int_{\mathrm{Y}^n} \varphi_n(x^n, y^n) d\nu^n(y^n) \\ &= \int_{\mathrm{Y}^n} \frac{p(y_1, y_2|x_1, x_2) \ldots p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(y_2|x_2) \ldots p(y_{n-1}|x_{n-1})} d\nu^n(y) \\ &= \int_{\mathrm{Y}^{n-1}} \left[\int_{\mathrm{Y}} \frac{p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(y_{n-1}|x_{n-1})} d\nu(y_n)\right] \varphi_{n-1}(x^{n-1}, y^{n-1}) d\nu^{n-1}(y) \\ &= \int_{\mathrm{Y}^{n-1}} \left[\frac{p(y_{n-1}|x_{n-1}, x_n)}{p(y_{n-1}|x_{n-1})}\right] \varphi_{n-1}(x^{n-1}, y^{n-1}) d\nu^{n-1}(y) \\ &= \int_{\mathrm{Y}^{n-1}} \varphi_{n-1}(x^{n-1}, y^{n-1}) d\nu^{n-1}(y) \end{aligned}$$

the last equality following from (H). So, after n steps we have $\int_{\mathrm{Y}^n} \varphi_n(x^n, y^n) d\nu^n(y^n) = 1$ which completes the proof. □

Let us notice that the following symmetrical condition $p(y_i|x_{i-1}, x_i) = p(y_i|x_i)$ can replace (H); the proof is modified by starting the integration from left instead of right. Conversely, let us study the reciprocal proposition in a particular case.

Proposition 2.3. *Let Z be a stationary PMC associated with X and Y and defined by $p(z_1, z_2)$. Each X_i takes its values in a finite set $\Omega = \{\omega_1, \ldots, \omega_k\}$, and each Y_i takes its values in Y. Furthermore, we assume that $p(y_i)$ is the density of the distribution of Y_i with respect to a measure ν on Y (in practice, Y is often R^m and ν the Lesbegue measure).*

Assuming the following hypothesis (H'): $p(y_2|x_1, x_2) = p(y_2|x_2, x_3)$ for each $(x_1, x_2, x_3, y_3) \in \Omega^3 \times \mathrm{Y}$ such that $x_1 = x_3$, the Markovianity of X implies $p(y_i|x_i, x_{i+1}) = p(y_i|x_i)$ for each $(x_i, x_{i+1}) \in \Omega^2$.

Proof. Let us consider the first three variables $(Z_1, Z_2, Z_3) = (X_1, Y_1, X_2, Y_2, X_3, Y_3)$. We have

$$p(z_1,z_2,z_3)=\frac{p(z_1,z_2)p(z_2,z_3)}{p(z_2)}=\frac{p(x_1,x_2)p(y_1,y_2|x_1,x_2)p(x_2,x_3)p(y_2,y_3|x_2,x_3)}{p(x_2)p(y_2|x_2)}=\left[\frac{p(x_1,x_2)p(x_2,x_3)}{p(x_2)}\right]\left[\frac{p(y_1,y_2|x_1,x_2)p(y_2,y_3|x_2,x_3)}{p(y_2|x_2)}\right]. \tag{2.3}$$

The sequence (X_1,X_2,X_3) being Markovian its distribution is $\frac{p(x_1,x_2)p(x_2,x_3)}{p(x_2)}$ and, thus, according to (2.3), the distribution of (Y_1,Y_2,Y_3) conditional on $(X_1,X_2,X_3)=(x_1,x_2,x_3)$ is

$$p(y_1,y_2,y_3|x_1,x_2,x_3)=\frac{p(y_1,y_2|x_1,x_2)p(y_2,y_3|x_2,x_3)}{p(y_2|x_2)}. \tag{2.4}$$

The integral of (2.4) with respect to y_1,y_2,y_3 is then equal to 1; after having integrated it with respect to y_1,y_3, we obtain

$$\int_Y \frac{p(y_2|x_1,x_2)p(y_2|x_2,x_3)}{p(y_2|x_2)}d\nu(y_2)=1. \tag{2.5}$$

Finally, the Markovianity of X implies (2.5). We will show that (2.5) implies $p(y_2|x_1,x_2)=p(y_2|x_2)$ by considering it as a scalar product. For a fixed x_2, let us put $g(y_2)=1/p(y_2|x_2)$. The function g is defined on $\mathrm{Y}^+\subset\mathrm{Y}$, such that $y_2\in\mathrm{Y}^+$ implies $p(y_2|x_2)>0$. Let us consider the set $L^2(g,\nu)$ of all functions $f:Y^+\to R$ such that $\int_{Y^+}g(y_2)f^2(y_2)d\nu(y_2)$ exists. As the function g is strictly positive, $\langle f_1,f_2\rangle=\int_{Y^+}g(y_2)f_1(y_2)f_2(y_2)d\nu(y_2)$ is a scalar product on $L^2(g,\nu)$. To simplify the notations, let us put $f_{x_1}(y_2)=p(y_2|x_1,x_2)$ and $f_{x_3}(y_2)=p(y_2|x_2,x_3)$, where x_2 is fixed. As all f_{x_1} (for each $x_1\in\Omega$) vanish outside Y^+ ($p(y_2|x_1,x_2)>0$ implies $p(y_2|x_2)>0$), they are equal if and only if $f^+_{x_1}$, their restrictions to Y^+, are equal. Now, we can see by taking $x_1=x_3$ in (2.5), that $f^+_{x_1}$ and $f^+_{x_3}$, which are equal because of the hypothesis (H'), are in $L^2(g,\nu)$. So, we have k vectors $f^+_{\omega_1},\ldots,f^+_{\omega_k}$ (which are restrictions of $f_{\omega_1},\ldots,f_{\omega_k}$ to Y^+) such that $\langle f^+_{\omega_i},f^+_{\omega_j}\rangle=1$ for each $1\le i,j\le k$ (which gives, in particular, that $\|f^+_{\omega_1}\|=\ldots=\|f^+_{\omega_k}\|=1$). This implies that they are all equal—and, thus, all $f_{\omega_1},\ldots,f_{\omega_k}$ are equal, which completes the proof. □

So, according to the Proposition 2.3, the classical hidden Markov chains, in which X is a Markov chain, cannot take into account the situations in which $p(y_{n-1}|x_{n-1},x_n)$ does depend on x_n. This can be a drawback in real situations in which such dependencies occur. For example, let us consider the problem of statistical image segmentation with two classes "forest" and "water:" $\mathrm{X}=\{F,W\}$. For $x_{n-1}=F$, the random variable Y_{n-1} models the natural variability of the forest and, possibly, other "noise" which is considered absent here. Considering $(x_{n-1},x_n)=(F,F)$ and $(x_{n-1},x_n)=(F,W)$ as two possibilities for (x_{n-1},x_n), it seems quite natural to consider that $p(y_{n-1}|F,F)$ and $p(y_{n-1}|F,W)$ can be different. In fact, in the second case, the trees are near water, which can make them greener or higher, say, giving them a different visual aspect. More generally, the possible dependence of $p(y_{n-1}|x_{n-1},x_n)$ on x_n allows one to model easily the fact that the visual aspect of a given class can be different near a boundary than inside a large set of pixels of a same class. So, a kind of "nonstationarity," which models the fact that the "noise" can be different close to "boundaries," can be taken into account in the frame of a "stationary" PMC model.

Let us briefly return to the classical model specified in Example 2.1 above. As we have $p(z_i,z_{i+1})=p(x_i,x_{i+1})p(y_i|x_i)p(y_{i+1}|x_{i+1})$, it is immediate to see that $p(y_{n-1}|x_{n-1},x_n)=p(y_{n-1}|x_{n-1})$ and, so, according to Proposition 2.1, we find again the fact that X is a Markov chain defined by $p(x_i,x_{i+1})$. Furthermore, we note that $p(x_{n-1}|y_{n-1},y_n)\neq p(x_{n-1}|y_{n-1})$, which is consistent with the well-known fact that Y is not a Markov chain.

Example 2.2. Let us complicate slightly the model specified in Example 2.1 above. Let $p(z_i,z_{i+1})=p(x_i,\,x_{i+1})p(y_i,\,y_{i+1}|x_i,\,x_{i+1})$, where $p(y_i,y_{i+1}|x_i,x_{i+1})$ are Gaussian distributions with nonzero correlations. Then, there are two possibilities:

1. $p(y_{n-1}|x_{n-1},x_n)=p(y_{n-1}|x_{n-1})$, which means that the mean and variance of $p(y_{n-1}|x_{n-1},x_n)$ do not depend on x_n. In this case, Proposition 2.2 is applicable and X is a Markov chain.
2. The mean or the variance of $p(y_{n-1}|x_{n-1},x_n)$ depends on x_n. In this case, Proposition 2.3 is applicable and X is not a Markov chain.

3 Pairwise Markov Chains and Hidden Markov Chains

The aim of this section is to specify some differences, considering a classical situation, that the greater generality of PMC with respect to HMC implies in practical calculus. So, let us consider a PMC given by $p(z)=p(x,y)=p(z_1)p(z_2|z_1)\ldots p(z_n|z_{n-1})$, which gives an HMC when

$$p(z_1)=p(x_1)p(y_1|x_1)\text{and }p(z_i|z_{i-1})=p(x_i|x_{i-1})p(y_i|x_i). \tag{3.1}$$

In HMC, many processing steps are based on the feasibility of recursive formulas for the so-called "forward" and "backward" probabilities. Let us consider:

$$\alpha^i(x^i)=p(y^1,\ldots,y^{i-1},z^i) \tag{3.2}$$

$$\beta^i(x^i)=p(y^{i+1},\ldots,y^n|z^i) \tag{3.3}$$

"forward" and "backward" probabilities associated with the PMC above. We can see that the forward probability (3.2) is the same as the HMC one, and that the backward probability gives the classical HMC one $\beta^i(x^i)=p(y^{i+1},\ldots,y^n|x^i)$ when (3.1) is verified. Furthermore, the PMC forward and backward probabilities can be recursively calculated by

$$\alpha^1(x^1)=p^y(z^1),\text{and }\alpha^{i+1}(x^{i+1})=\sum_{x^i\in\mathrm{X}^{Ni}}\alpha^i(x^i)p(z^{i+1}|z^i)\text{for }2\le i\le n \tag{3.4}$$

$$\beta^n(x^n)=1,\text{and }\beta^i(x^i)=\sum_{x^{i+1}\in\mathrm{X}^{Ni+1}}\beta^{i+1}(x^{i+1})p(z^{i+1}|z^i)\text{for }1\le i\le n-1 \tag{3.5}$$

which gives the classical calculations when (3.1) is verified. Otherwise, we have

$$p(x_i,x_{i+1}|y)=\frac{\alpha^i(x_i)p(z_{i+1}|z_i)\beta^{i+1}(x_{i+1})}{\sum_{(\omega_1,\omega_2)\in\Omega^2}\alpha^i(\omega_1)p(y_{i+1},\omega_2|y_i,\omega_1)\beta^{i+1}(\omega_2)} \tag{3.6}$$

which gives the classical HMC formula when (3.1) is verified. Of course, (3.6) gives $p(x_1|y)$ and $p(x_{i+1}|\,x_i,y)$, which defines the Markov distribution of X conditional to $Y=y$. Furthermore, we have the same formula (3.7) as in the HMC case, which allows us to calculate the Bayesian MPM restoration of the hidden X:

$$p(x_i|y)=\frac{\alpha^i(x^i)\beta^i(x^i)}{\sum_{\omega\in\Omega}\alpha^i(\omega)\beta^i(\omega)}. \tag{3.7}$$

To better situate PMC with respect to HMC, let us examine, by considering a simple case, how the number of parameters grows when generalizing HMC to PMC. Consider the case of two classes $\mathrm{X}=\{\omega_1,\omega_2\}$, and assume that the distribution of Y conditional on

$X = x$ is Gaussian. In the classical HMC considered above, we have $p(z_i, z_{i+1}) = p(x_i, x_{i+1})p(y_i|x_i)p(y_{i+1}|x_{i+1})$, with $p(y_i|x_i)$ and $p(y_{i+1}|x_{i+1})$ Gaussian distributions on R. Thus, we have four parameters for $p(x_i, x_{i+1})$ and four parameters (two means and two variances) for the two distributions $p(y_i|\omega_1)$ and $p(y_i|\omega_2)$ (which do not depend on i). In the PMC case, we have $p(z_i, z_{i+1}) = p(x_i, x_{i+1})p(y_i, y_{i+1}|x_i, x_{i+1})$, with $p(y_i, y_{i+1}|x_i, x_{i+1})$ Gaussian distributions on R^2. As above, we have four parameters for $p(x_i, x_{i+1})$ but the number of "noise" parameters is greater. In fact, we have four Gaussian distributions on R^2, which gives 20 parameters (two means, two variances, and one covariance for each of them). So, we have eight parameters in the HMC case and 24 in the PMC one. How does one estimate all these parameters? When both X and Y are observed, one can use the classical "empirical" estimators. Denoting by

$$\hat{m}_{kj} = \begin{pmatrix} \hat{m}^1_{kj} \\ \hat{m}^2_{kj} \end{pmatrix}$$

the empirical estimate of the mean vector of the Gaussian distribution $p(y_i, y_{i+1}|\omega_k, \omega_j)$, and by $\hat{\Gamma}_{kj}$ the empirical estimate its variance-covariance matrix, we have:

$$\hat{p}(\omega_k, \omega_j) = \frac{1}{n-1}\sum_{i=1}^{n-1} 1_{[(x_i, x_{i+1})=(\omega_k, \omega_j)]}, \tag{3.8}$$

$$\begin{pmatrix} \hat{m}^1_{kj} \\ \hat{m}^2_{kj} \end{pmatrix} = \left[\sum_{i=1}^{n-1} \begin{pmatrix} y_i \\ y_{i+1} \end{pmatrix} 1_{[(x_i, x_{i+1})=(\omega_k, \omega_j)]}\right] \Big/ \left[\sum_{i=1}^{n-1} 1_{[(x_i, x_{i+1})=(\omega_k, \omega_j)]}\right], \tag{3.9}$$

$$\hat{\Gamma}_{kj} = \left[\sum_{i=1}^{n-1} \begin{pmatrix} y_i - \hat{m}^1_{kj} \\ y_{i+1} - \hat{m}^2_{kj} \end{pmatrix}^t \begin{pmatrix} y_i - \hat{m}^1_{kj} \\ y_{i+1} - \hat{m}^2_{kj} \end{pmatrix} 1_{[(x_i, x_{i+1})=(\omega_k, \omega_j)]}\right] \Big/ \left[\sum_{i=1}^{n-1} 1_{[(x_i, x_{i+1})=(\omega_k, \omega_j)]}\right]. \tag{3.10}$$

When only Y is observed, (3.8), (3.9), and (3.10) can be used in the ICE method, as treated in the next section.

4 Parameter Estimation from Incomplete Data

In this section, we very briefly mention how the general Iterative Conditional Estimation (ICE) can be used to learn a PMC from Y; see [15], [20] for further description of the tools used. Let Z be a stationary PMC defined by a density $p(z_1, z_2)$ depending on a parameter $\theta \in \Theta$. The problem is to estimate θ from a sample $y = (y_1, \ldots, y_n)$. The use of ICE is possible once

i. there exists an estimator of θ from the complete data: $\hat{\theta} = \hat{\theta}(z) = \hat{\theta}((x_1, y_1), \ldots, (x_n, y_n))$;
ii. for each $\theta \in \Theta$, either the conditional expectation $E_\theta[\hat{\theta}(Z)|Y = y]$ is computable, or simulations of X according to its distribution conditional to $Y = y$ are feasible.

ICE produces a sequence of parameters $(\theta^q)_{q \in N}$ in the following way:

1. Initialize $\theta = \theta^0$;
2. for $q \in N$,
   a. set $\theta^{q+1} = E_{\theta^q}[\hat{\theta}(Z)|Y = y]$ if the conditional expectation is computable;
   b. if not, simulate l realizations $x^1, \ldots, x^l$ of X according to its distribution conditional to $Y = y$ and based on θ^q, and set $\theta^{q+1} = \frac{\hat{\theta}(x^1, y) + \ldots + \hat{\theta}(x^l, y)}{l}$.

We note that, in general, θ is a vector and thus it may happen that some of its components are re-estimated by the conditional expectation 2a, and the remaining ones are re-estimated using simulations 2b. Otherwise, we observe that (i) is a very weak assertion; in fact, being able to estimate θ from X and Y is the minimum we require for estimating θ from Y alone.

ICE is easily applicable in the Gaussian PMC described in the previous section. The parameter θ includes here the probabilities $p(x_i, x_{i+1})$ (so, k^2 real parameters for k classes), and the means and the covariance matrices of the k^2 Gaussian distributions $p(y_1, y_2|x_1, x_2)$ on R^2. We then see that (i) is verified with $\hat{\theta}$ given by (3.8), (3.9), and (3.10) and, X being a Markov chain conditionally to $Y = y$, its samplings are feasible, which gives (ii). More precisely, the conditional expectation 2a is calculable when the parameters $p(\omega_k, \omega_j)$ are concerned, which gives

$$\begin{aligned} p^{q+1}(\omega_k, \omega_j) &= E_{\theta^q}[\hat{p}(\omega_k, \omega_j)(X)|Y = y] \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1} p^q[(x_i, x_{i+1}) = (\omega_k, \omega_j)|y], \end{aligned}$$

with $p^q[(x_i, x_{i+1}) = (\omega_k, \omega_j)|y]$ given by (3.6). On the contrary, the conditional expectation of the estimates (3.9) and (3.10) is not computable and, so, to re-estimate the noise parameters, we use the sampled realizations $x^1, \ldots, x^l$ of X, as specified in 2b, with $\hat{m}_{kj}$ and $\hat{\Gamma}_{kj}$ given by (3.9) and (3.10).

Let us briefly mention how the "generalized mixture" estimation method in the case of correlated and nonnecessarily Gaussian sensors proposed in [20] can be adapted to the PMC case. As above, for k classes, $p(z_1, z_2) = p(x_1, x_2)p((y_1, y_2)|(x_1, x_2))$ is defined by k^2 parameters $p(x_1, x_2)$ and k^2 probability densities $p((y_1, y_2)|(x_1, x_2))$ on R^2, which will be denoted by f_{ij}, with $1 \le i, j \le k$. So, the distribution of each random vector $Y^*_m = (Y_{2m-1}, Y_{2m})$ is a mixture of k^2 distributions on R^2, and we dispose of a sample $y^*_1 = (y_1, y_2), \ldots, y^*_n = (y_{2n-1}, y_{2n})$ distributed according to this mixture (the number of observations is assumed to be even). We may then apply the "generalized" ICE (ICE-GEMI) described in [20], which allows one to search the k^2 densities f_{ij} in different general sets of densities, possibly including correlated and non-Gaussian components.

5 Experiments

In this section, we present two series of results: the first concerns data simulated according to a PMC and the second concerns a two classes image corrupted by correlated Gaussian noise.

So, let Z be a stationary PMC, with $\Omega = \{\omega_1, \omega_2\}$ and $Y = R$. The distribution of Z is defined by $p(z_1, z_2) = p(x_1, x_2)p((y_1, y_2)|(x_1, x_2))$; so, we have to choose $p(\omega_1, \omega_1)$, $p(\omega_1, \omega_2)$ $p(\omega_2, \omega_1)$, $p(\omega_2, \omega_2)$, and four Gaussian distributions $p(y_1, y_2|\omega_1, \omega_1)$, $p(y_1, y_2|\omega_1, \omega_2)$, $p(y_1, y_2|\omega_2, \omega_1)$, and $p(y_1, y_2|\omega_2, \omega_2)$ on R^2. The parameters of the latter four Gaussian distributions are specified in Table 1, and we consider two cases ($p(\omega_1, \omega_1) = p(\omega_2, \omega_2) = 0.48$, $p(\omega_1, \omega_2) = p(\omega_2, \omega_1) = 0.01$ in case 1, and $p(\omega_1, \omega_1) = p(\omega_2, \omega_2) = p(\omega_2, \omega_1) = p(\omega_1, \omega_2) = 0.25$ in case 2) for the four distributions $p(\omega_i, \omega_j)$.

The PMC (X, Y) is then sampled, giving $(X, Y) = (x_r, y)$ (we put x_r to specify that it is the "real" realization of X). The realization of X is then estimated by the Bayesian MPM method using the real PMC model, which gives $\hat{x}_{PMC}$, and by the Bayesian MPM method using a HMC, which gives $\hat{x}_{HMC}$. Comparing $\hat{x}_{PMC}$ and $\hat{x}_{HMC}$ to x_r gives then the error ratios τ_{PMC} and τ_{HMC}. The HMC used can be seen as an "approximation" of the PMC; its distribution is given by the same $p(\omega_1, \omega_1), p(\omega_2, \omega_2), p(\omega_1, \omega_2), p(\omega_2, \omega_1)$, and by two Gaussian densities on R (recall that $p(y_i, y_{i+1}|x_i, x_{i+1}) = p(y_i|x_i)p(y_{i+1}|x_{i+1})$), which are of mean 0 and standard deviation 14 for $p(y_i|x_i = \omega_1)$, and of mean 10 and standard deviation 20 for $p(y_i|x_i = \omega_2)$.

The length of the simulated chains is $n = 4,000$, and the error ratios presented are the means of the error ratio obtained with 250 independent experiments.

TABLE 1
Parameters of Four Gaussian Distributions

		Parameters of Gaussian distributions $p(y_i, y_{i+1}\|x_i, x_{i+1})$					Case 1		Case 2	
x_i	x_{i+1}	μ_1	μ_2	σ_1	σ_2	ρ				
ω_1	ω_1	0	0	14	14	0.9	τ_{HMC}	τ_{PMC}	τ_{HMC}	τ_{PMC}
ω_1	ω_2	2	8	14	20	0.1				
ω_2	ω_1	8	2	20	14	0.1				
ω_2	ω_2	10	10	20	20	0.9	15.0%	12.2%	36.9%	12.5%

Parameters of four Gaussian distributions $p(y_i, y_{i+1}|x_i, x_{i+1})$, *where* $p(x_i, x_{i+1})$ *is given by* $p(\omega_1, \omega_1) = p(\omega_2, \omega_2) = 0.48$ *in case 1, and by* $p(\omega_1, \omega_1) = p(\omega_2, \omega_2) = p(\omega_2, \omega_1) = p(\omega_1, \omega_2) = 0.25$ *in case 2.* τ_{PMC} *and* τ_{HMC} *are the error ratios of MPM restoration based on PMC and HMC, respectively.*

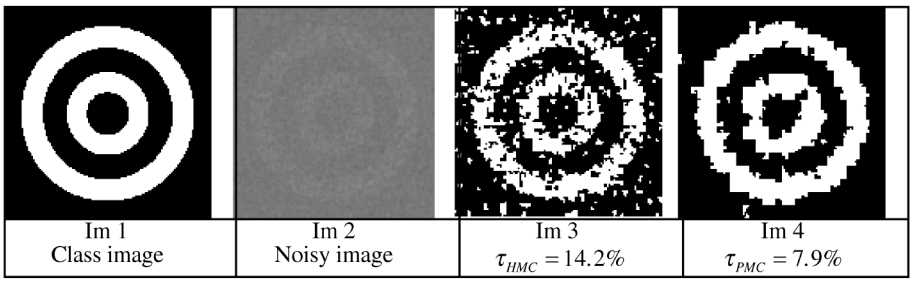

Fig. 1. Class image, its noisy version, and two unsupervised segmentations based on HMC and PMC. Parameters estimated with ICE. τ is the error ratio.

Of course, the fact that τ_{PMC} is smaller than τ_{HMC} is not surprising because it is in accordance to the very Bayesian theory. However, the results obtained show that in some situations, as in case 2 in Table 1, the difference of error ratios may be quite significant.

The interest of the second series of experiments below is double. First, the results presented show that PMC-based unsupervised segmentation methods can be more efficient than the HMC-based ones. Second, the results obtained allow us to propose some answer to the following robustness problem: when the data neither suit a PMC nor a HMC model, does the PMC-based unsupervised MPM restoration method work better than the HMC-based one?

Let us consider a two classes image (Im 1, Fig. 1), corrupted with correlated noise (Im 2, Fig. 1). More specifically, the observed field is $Y_s = \sigma_{x_s} W_s + \mu_{x_s} + a \sum_{i=1}^{4} (\sigma_{x_{s_i}} W_{s_i} + \mu_{x_{s_i}})$, where $W = (W_s)$ is a white Gaussian noise with variance 1, $s_1, \ldots, s_4$ are four neighbors of s, and $x_s = \omega_1$ (white) or $x_s = \omega_2$ (black). The set of pixels, which is here of size $n = 128 \times 128$, is then transformed in a sequence $s_1, \ldots, s_n$ via the Hilbert-Peano scan, as described in Fig. 2 (see also [15]). Putting $X_i = X_{s_i}$ and $Y_i = Y_{s_i}$, we consider that Im 1 is a realization of $X = (X_1, \ldots, X_n)$, and Im 2 is a realization of $Y = (Y_1, \ldots, Y_n)$. Of course, the distribution of the pairwise process $Z = (X, Y)$ is very complicated and, in particular, we can see that $p(y|x)$ is not necessarily a Markov distribution. So, Z is neither a PMC nor a HMC. The question is then to know whether the use of PMC instead of HMC in such situations can improve the segmentation results. So, Z is considered as a HMC on the one hand, and as a PMC, on the other hand. In both cases, the corresponding MPM restorations are then performed in an unsupervised manner, the parameters being estimated by ICE accordingly to the description in Section 4 (in the cases considered, ICE turns out to be little sensitive to the initialization of the parameter values).

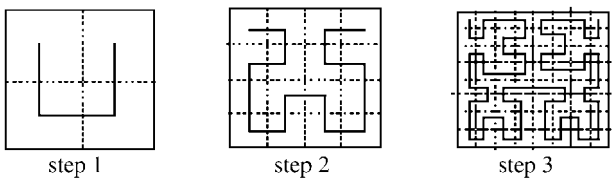

Fig. 2. Construction of the Hilbert-Peano scan.

We have performed numerous simulations, with different parameters a, μ_{x_s}, σ_{x_s}, and it turns out that PMC-based MPM works consistently better than the HMC-based one. The results of one series of simulations, corresponding to $a = 0.4$, $\mu_{\omega_1} = 120$, $\mu_{\omega_2} = 125$, $\sigma_{\omega_1} = 50$, and $\sigma_{\omega_2} = 75$, are presented in Fig. 1. Otherwise, the ICE-based estimation method used behaves very well; in fact, using the parameters estimated by formulas (3.15), (3.16), and (3.17) from complete data provides nearly the same error ratio of $\tau_{PMC} = 8.0\%$.

6 Conclusions and Perspectives

We proposed in this paper a new Pairwise Markov Chain (PMC) model. Having an unobservable process $X = (X_1, \ldots, X_n)$ and an observed process $Y = (Y_1, \ldots, Y_n)$, the idea was to consider the Markovianity of the couple $Z = (X, Y)$. This idea is analogous to that having lead to the Pairwise Markov Random Field (PMRF) model recently proposed in [21], although significant differences between the two models exist.

We have discussed some advantages of the new model with respect to the classical Hidden Markov Chain (HMC) model, which appears as a particular case. In particular, we gave a necessary condition for the hidden process to be a Markov one, which shows that the classical HMC cannot model some intuitively interesting cases.

A parameter estimation method, based on the general Iterative Conditional Estimation (ICE) procedure, has been presented and some extensions, valid when the exact nature of the noise is not known [9], [15], [20], have been briefly mentioned. Finally, the interest of PMC has been validated by some preliminary experiments.

As perspectives, let us mention the possibility of using HMC models in multiresolution image segmentation problems [17], which could thus be generalized to some PMC models. More generally, HMRF and HMC are particular cases of Hidden Markov models on networks [7]. So, different generalizations of PMC proposed here—and PMRF proposed in [21]—to Pairwise Markov Processes on Networks could undoubtedly be considered and be of interest in some situations.

ACKNOWLEDGMENTS

The author would like to thank Alain Hillion, from Ecole Nationale Supérieure des Télécommunications de Bretagne, and Francois Desbouvries, from Institut National des Télécommunications, for numerous discussions that greatly helped in the writing of this paper. He would also like to thank Stéphane Derrode, from Ecole Nationale Supérieure de Physique de Marseille, who performed the simulations.

REFERENCES

[1] K. Aas, L. Eikvil, and R.B. Huseby, "Applications of Hidden Markov Chains in Image Analysis," *Pattern Recognition,* vol. 32, no. 4, pp. 703-713, 1999.

[2] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statstical Analysis of Probabilistic Functions of Markov Chains," *Ann. Math. Statistics,* vol. 41, pp. 164-171, 1970.

[3] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. Royal Statistical Soc., Series B,* vol. 48, pp. 259-302, 1986.

[4] M.-Y. Chen, A. Kundu, and J. Zhou, " Off-Line Handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 5, pp. 481-496, May 1994.

[5] J.-L. Chen and A. Kundu, "Unsupervised Texture Segmentation Using Multi-Channel Decomposition and Hidden Markov Model," *IEEE Trans. Image Processing,* vol. 4, no. 5, pp. 603-619, 1995.

[6] G.A. Churchill, "Hidden Markov Chains and the Analysis of Genome Structure," *Computers and Chemistry,* vol. 16, no. 2, pp. 107-115, 1992.

[7] R.G. Cowell, A.P. David, S.L. Lauritzen, and D.J. Spiegelhalter, *Probabilistic Networks and Expert Systems.* New York: Springer-Verlag, 1999.

[8] J. Dai, "Hybrid Approach to Speech Recognition Using Hidden Markov Models and Markov Chains," *IEE Proc. Vision, Image, and Signal Processing,* vol. 141, no. 5, pp. 273-279, 1994.

[9] Y. Delignon, A. Marzouki, and W. Pieczynski, "Estimation of Generalized Mixture and Its Application in Image Segmentation," *IEEE Trans. Image Processing,* vol. 6, no. 10, pp. 1364-1375, 1997.

[10] J.-P. Delmas, "An Equivalence of the EM and ICE Algorithm for Exponential Family," *IEEE Trans. Signal Processing,* vol. 45, no. 10, pp. 2613-2615, 1997.

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc., Series B,* vol. 39, pp. 1-38, 1977.

[12] A. El-Jacoubi, M. Gilloux, R. Sabourin, and C.Y. Suen, "An HMM-Based Approach for Off-Line Unconstrained Handwritten Word Modeling and Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 8, pp. 752-760, Aug. 1999.

[13] G.D. Fornay, "The Viterbi Algorithm," *Proc. IEEE,* vol. 61, no. 3, pp. 268-277, 1973.

[14] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, pp. 721-741, 1984.

[15] N. Giordana and W. Pieczynski, "Estimation of Generalized Multisensor Hidden Markov Chains and Unsupervised Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 5, pp. 465-475, May 1997.

[16] G.K. Kaleh and R. Vallet, "Joint Parameter Estimation and Symbol Detection for Linear or Nonlinear Unknown Channels," *IEEE Trans. Comm.,* vol. 42, no. 7, pp. 2406-2413, July 1994.

[17] J.-M. Laferté, P. Pérez, and F. Heitz, "Discrete Markov Image Modeling and Inference on the Quadtree," *IEEE Trans. Image Processing,* vol. 9, no. 3, pp. 390-404, 2000.

[18] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic Solution of Ill-Posed Problems in Computational Vision," *J. Am. Statistical Assoc.,* vol. 82, pp. 76-89, 1987.

[19] M. Mignotte, C. Collet, P. Pérez, and P. Bouthémy, "Sonar Image Segmentation Using an Unsupervised Hierarchical MRF Model," *IEEE Trans. Information Processing,* vol. 9, no. 7, pp. 1216-1231, 2000.

[20] W. Pieczynski, J. Bouvrais, and C. Michel, "Estimation of Generalized Mixture in the Case of Correlated Sensors," *IEEE Trans. Image Processing,* vol. 9, no. 2, pp. 308-311, 2000.

[21] W. Pieczynski and A.-N. Tebbache, "Pairwise Markov Random Fields and Segmentation of Textured Images," *Machine Graphics & Vision,* vol. 9, no. 3, pp. 705-718, 2000.

[22] W. Qian and D.M. Titterington, "On the Use of Gibbs Markov Chain Models in the Analysis of Images Based on Second-Order Pairwise Interactive Distributions," *J. Applied Statistics,* vol. 16, no. 2, pp. 267-282, 1989.

[23] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE,* vol. 77, no. 2, pp. 257-286, 1989.

[24] C. Raphael, "Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 4, pp. 360-370, 1999.

[25] F. Salzenstein and W. Pieczynski, "Sur le Choix de Méthode deSegmentation Statistique d'Images," *Traitement du Signal,* vol. 15, no. 2, pp. 119-128, 1998.

[26] A.D. Wilson and A.F. Bobick, "Parametric Hidden Markov Models for Gesture Recognition," *IEEE Trans. Image Processing,* vol. 8, no. 9, pp. 884-900, 1999.