Pairwise Markov Chains and Bayesian Unsupervised Fusion

Wojciech Pieczynski
Département Signal et Image
Institut National des Télécommunications
9, rue Charles Fourier, 91000 Evry, France

Abstract - *We propose a new model called a Pairwise Markov Chain (PMC), which generalises the classical Hidden Markov Chain (HMC) model. The PMC model is more general than HMC in that the process one wants to estimate 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 speech recognition, multisensor image segmentation, among others. Furthermore, we propose a new method of parameter estimation, which allows one to perform unsupervised restoration with PMC. The method proposed is valid even with non Gaussian and possibly correlated noise. Furthermore, the very form of the statistical distribution of the noise need not be known exactly. All that is required is that for each class the form of the noise distribution belongs to a given set of forms.*

Keywords: Markov chain, hidden data, Bayesian restoration, unsupervised segmentation, iterative conditional estimation, pairwise Markov chain.

1. Introduction

The field of applications of Hidden Markov Models is extremely wide. Among this family of models, Hidden Markov Chains (HMC) are among the most frequently used. In pattern recognition and image processing area, HMC can be used in image segmentation [6, 21, 24], hand-written word recognition [7], acoustic musical signal recognition [23], or even gesture recognition [26]. Some other areas of possible application are speech recognition [22] and communications [12]. Multisensor images, or even multisensor and multiresolution images, can still be segmented using hidden Markov chains [11, 8]. *A priori*, Hidden Markov Random Fields (HMRF) are better suited to deal with the image segmentation problem [2, 4, 10, 13, 15], although, HMC based segmentation methods can be competitive in some particular situations [24], and they are much faster that the HMRF based ones. The success of such models is due to the fact that when the unobservable, or hidden, signal can be modelled 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) [1, 6, 9]. These restoration methods use the distribution of the hidden process conditional to the observations, which is called its "posterior" distribution. This posterior distribution can then be seen as a fusion of the information contained in the observation with the "prior" information, which is modelled by the "prior" distribution of the hidden process. Furthermore, such restoration techniques can be rendered unsupervised by applying some parameter estimation method, like Expectation-Maximization (EM) [1, 5, 12, 24], or Iterative Conditional Estimation (ICE) [17, 19, 24]. For instance, let us consider the following hidden Markov model: $X=(X_1,...,X_n)$ is a Markov chain, with all X_i taking their values in the set of classes $\Omega=\{\omega_1,...,\omega_k\}$, and $Y=(Y_1,...,Y_n)$ 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 $P_X^{Y=y}$ be still a Markov distribution. We shall insist that the Markovianity of $P_X^{Y=y}$ is *essential* to applying the 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 some simple distributions $P_Y^{X=x}$ such that when P_X is a Markov distribution, $P_X^{Y=y}$ is no longer a Markovian one. The aim of this paper is to propose a model which would be more general that the hidden Markov chain model and in which the posterior distribution $P_X^{Y=y}$ would *always* be a Markov chain distribution. 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. Let us notice that for a given $P_Y^{X=x}$ we can consider a HMC in which $P_X^{Y=y}$ is not Markovian, and a PMC in which $P_X^{Y=y}$ is Markovian and P_X is not.

Me may remark that having P_X not necessarily Markovian could be seen as a drawback; indeed, this distribution models our prior information about the process. This is undoubtedly a drawback in the situations in which we effectively know that X is a Markov chain and we know all parameters defining P_X. However, in numerous real situations the distribution P_X is not known exactly and has to be estimated. So, the only "a priori" information is that X is a Markov chain and its parameters have to be estimated from Y. So, this seems to be a little bit inconsistent because a part of the "a priori" information is estimated from Y, and thus becomes "a posteriori" information. Furthermore, in many real situations the Markovianity of X is only assumed and is not strictly established. Finally, we can say that when the Markovian nature of X is not sure but only assumed (with unknown parameters), the use of a PMC model is not necessarily less relevant that the use a HMC model.

The Bayesian restoration methods based on the PMC model can be rendered unsupervised by considering some model parameter estimation methods using only the observed data Y. We propose an original method which allows one to deal with the cases when the noise in not necessarily independent or Gaussian (although noise is often assumed Gaussian, the non Gaussian case is of interest in numerous situations [4, 11, 14, 16]). Furthermore, one can not know the exact forms of the noise distributions, and these forms can vary with the class. However, one has to know a set of possible forms. Although original, our method is inspired from the method recently proposed in [18], which gave acceptable numerical results.

The PMC model studied in this paper draws from the same idea as the Pairwise Markov Random Field (PMRF) model introduced in [20]. However, it is well known that Markov chains and Markov fields are different models and so the different properties of PMC discussed below are not necessarily true in the PMRF case.

The paper is organized as follows.

The PMC model is introduced in next section and its different properties are contrasted to those of the classical HMC model. An original PMC parameter estimation method, called PMC Iterative Conditional Estimation (PMC-ICE) is then described. Conclusions and perspectives are in third section.

2. Pairwise Markov chains

2.1 Model

Let us consider two sequences of random variables $X=(X_1,\dots,X_n)$, and $Y=(Y_1,\dots,Y_n)$. 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 $Z=(Z_1,\dots,Z_n)$ 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. In order to simplify things 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 he distribution of X_{i+1} conditional on $X_i=x_i$, $p(y|x)$ will be the density of the distribution of Y conditional on $X=x$, and so on. 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 is defined by

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

where $p(.)$ are probability densities with respect to some measures.

Thus the distribution of a pairwise Markov chain is given by the densities $p(z_1,z_2)$, ..., $p(z_{n-1},z_n)$.

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

Proof.

Let us 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 :

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

1. We have

$$p(z_{i+1}|z_1,z_2,\ldots,z_i)=\frac{p(z_1,z_2,\ldots,z_i,z_{i+1})}{p(z_1,z_2,\ldots,z_i)}=$$

$$=\frac{\dfrac{p(z_1,z_2)\ldots p(z_i,z_{i+1})}{p(z_2)\ldots p(z_i)}}{\dfrac{p(z_1,z_2)p(z_2,z_3)\ldots p(z_{i-1},z_i)}{p(z_2)p(z_3)\ldots p(z_{i-1})}}=\frac{p(z_i,z_{i+1})}{p(z_i)}=p(z_{i+1}|z_i)$$

where we have applied, in the second equality, (2.2) with $m=1$ and $l=i+1$.

2. Let us put $p(y|x)=p^x(y)$. We have to show that $p^x(y_{m+1}|y_1,\ldots,y_m)=p^x(y_{m+1}|y_m)$.

We have

$$p^x(y_{m+1}|y_1,\ldots,y_m)$$

$$=p^x(y_1,\ldots,y_m,y_{m+1})\,/\,p^x(y_1,\ldots,y_m)$$

$$=\left[\frac{p(x,y_1,\ldots,y_m,y_{m+1})}{p(x)}\right]\Big/\left[\frac{p(x,y_1,\ldots,y_m)}{p(x)}\right]$$

$$=\frac{p(x,y_1,\ldots,y_m,y_{m+1})}{p(x,y_1,\ldots,y_m)}$$

$$=\frac{\int_{Y^{n-m-1}}p(x,y_1,\ldots,y_m,y_{m+1},y_{m+2},\ldots,y_n)dy_{m+2}dy_n}{\int_{Y^{n-m}}p(x,y_1,\ldots,y_m,y_{m+1},y_{m+2},\ldots,y_n)dy_{m+1}dy_{m+2}\ldots dy_n}$$

$$=\frac{p(z_1,z_2)\ldots p(z_m,z_{m+1})}{p(z_1,z_2)\ldots p(z_{m-1},z_m)}$$

$$\frac{\int_{Y^{n-m-1}}p(x_{m+2},\ldots,x_n,y_{m+1},y_{m+2},\ldots,y_n)dy_{m+2}dy_n}{\int_{Y^{n-m}}p(x_m,x_{m+1}\ldots,x_n,y_m,y_{m+1},y_{m+2},\ldots,y_n)dy_{m+1}dy_{m+2}dy_n}$$

$$=\frac{p(z_m,z_{m+1})\int_{Y^{n-m-1}}p(x_{m+1},\ldots,x_n,y_{m+1},y_{m+2},\ldots,y_n)dy_{m+2}dy_n}{\int_{Y^{n-m}}p(x_m,x_{m+1}\ldots,x_n,y_m,y_{m+1},y_{m+2},\ldots,y_n)dy_{m+1}dy_{m+2}dy_n}$$

$$=\frac{p(z_m,z_{m+1})\int_{Y^{n-m-1}}p(z_{m+1},\ldots z_n)\,dy_{m+2}dy_n}{\int_{Y^{n-m}}p(x_m,z_{m+1},\ldots z_n)\,dy_{m+1}dy_{m+2}dy_n}$$

$$=a(x_m,x_{m+1},y_m,y_{m+1},\ldots,y_n)$$

we notice that there is no $y_1,\ldots,y_{m-1}$ in $a(x_m,x_{m+1},y_m,y_{m+1},\ldots,y_n)$ (which is a transition matrix term), which completes the proof.

We can notice that the proof is quite analogous to the proof in the classical hidden Markov case: the terms of the transition matrix are obtained by a "backward" recursion.

3. See (2.2).

Example 2.1

Let us consider a classical hidden Markov chain with independent noise: $X=\{\omega_1,\ldots,\omega_k\}$ is a finite set of classes, X a classical Markov chain on X, and $Y=R$ 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

$$p(z)=p(x,y)=$$

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

If the chain X is homogeneous and stationary, we have $p(x_{i+1}|x_i)=\dfrac{p(x_i,x_{i+1})}{p(x_i)}$ 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 $X^n\times Y^n$. 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 sub-vectors, are concerned different p simply are probabilities, and, when the vector y, or some of its sub-vectors, are concerned, different p are densities with respect to the Lebesgue measure. When both are concerned, as in (2.2), p is a density with respect to a product of some counting measure with some Lebesgue measure.

Proposition 2.2

Let Z be a stationary PMC associated with X and Y and defined by $p(z_1,z_2)$.

If

(H) $p(y_i|x_i, x_{i+1}) = p(y_i|x_i)$

Then X is a Markov chain .Furthermore, the distribution of the Markov chain X is

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

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

Proof

As $p(x)$ is a marginal distribution of $p(x,y)$:

$$p(x) = \int_{Y^n} p(x,y)dv^n(y) = p(x)\int_{Y^n} p(y|x)dv^n(y)$$

so, if $p(x)$ can be written $p(x) = q(x)\varphi(x,y)$ with $\int_{Y^n}\varphi(x,y)dv^n(y) = 1$, $q(x)$ is necessarily $p(x)$.
We have :

$$p(x,y) = p(z) = \frac{p(z_1, z_2)\dots p(z_{n-1}, z_n)}{p(z_2)\dots p(z_{n-1})}$$

$$= \frac{\dfrac{p(z_1, z_2)p(x_1, x_2)\dots p(z_{n-1}, z_n)p(x_{n-1}, x_n)}{p(x_1, x_2)\dots p(x_{n-1}, x_n)}}{\dfrac{p(z_2)p(x_2)\dots p(z_{n-1})p(x_{n-1})}{p(x_2)\dots p(x_{n-1})}}$$

$$= \left[\frac{p(x_1, x_2)\dots p(x_{n-1}, x_n)}{p(x_2)\dots p(x_{n-1})}\right]\left[\frac{\dfrac{p(z_1, z_2)}{p(x_1, x_2)}\dots\dfrac{p(z_{n-1}, z_n)}{p(x_{n-1}, x_n)}}{\dfrac{p(z_2)\dots p(z_{n-1})}{p(x_2)\dots p(x_{n-1})}}\right]$$

$$= \left[\frac{p(x_1, x_2)\dots p(x_{n-1}, x_n)}{p(x_2)\dots p(x_{n-1})}\right]$$

$$\left[\frac{p(y_1, y_2|x_1, x_2)\dots p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(y_2|x_2)\dots p(y_{n-1}|x_{n-1})}\right] = q(x)\varphi(x,y)$$

It remains to show that $\int_{Y^n}\varphi(x,y)dv^n(y) = 1$. We have

$$\int_{Y^n}\varphi(x,y)dv^n(y)$$

$$= \int_{Y^n}\frac{p(y_1, y_2|x_1, x_2)\dots p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(y_2|x_2)\dots p(y_{n-1}|x_{n-1})}dv^n(y)$$

$$= \int_{Y^{n-1}}\left[\int_Y \frac{p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(y_{n-1}|x_{n-1})}dv(y_n)\right]$$

$$\left[\frac{p(y_1, y_2|x_1, x_2)\dots p(y_{n-2}, y_{n-1}|x_{n-2}, x_{n-1})}{p(y_2|x_2)\dots p(y_{n-2}|x_{n-2})}\right]dv^{n-1}(y)$$

$$= \int_{Y^{n-1}}\frac{p(y_1, y_2|x_1, x_2)\dots p(y_{n-2}, y_{n-1}|x_{n-2}, x_{n-1})}{p(y_2|x_2)\dots p(y_{n-2}|x_{n-2})}dv^{n-1}(y)$$

where the last equality follows from

$$\int_Y \frac{p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(y_{n-1}|x_{n-1})}dv(y_n)$$

$$= \frac{1}{p(y_{n-1}|x_{n-1})}\int_Y \frac{p(y_{n-1}, y_n|x_{n-1}, x_n)}{p(y_{n-1}|x_{n-1})}dv(y_n)$$

$$= \frac{p(y_{n-1}|x_{n-1}, x_n)}{p(y_{n-1}|x_{n-1})} = \frac{p(y_{n-1}|x_{n-1})}{p(y_{n-1}|x_{n-1})} = 1$$

which comes from $p(y_{n-1}|x_{n-1}, x_n) = p(y_{n-1}|x_{n-1})$.

So, after n we have $\int_{Y^n}\varphi(x,y)dv^n(y) = 1$.

Example 2.2

Let us 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 2.3

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 non null 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 thit case the 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 thit case, X is not a Markov chain.

So we can see how pairwise Markov chains generalise hidden Markov chains. In fact, as X is not a Markov chain in the the second case, the model is not a classical Hidden Markov Chain.
Let us illustrate the possible dependence of $p(y_{n-1}|x_{n-1}, x_n)$ on x_n in real situations by an example. Let us consider the problem of statistical image segmentation with two classes "forest" and "water" : $X = \{F, W\}$. 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.

2.2 Parameter estimation

In many applications in signal or image processing it is useful to dispose of unsupervised methods, which means that all model parameters are estimated, in a previous step, from the sole observation. We propose in this section an original method based on the general Iterative Conditional Estimation for Generalized Mixtures (ICE-GEMI) method. Such a method has been successfully applied in the case of non Gaussian and possibly correlated sensors [18]. As we shall see in the following, it is still possible to apply such kind of method in PMC, mainly because of the fact that the distribution of (Z_i, Z_{i+1}), which is a marginal distribution of Z, is given by the density $p(z_i, z_{i+1})$ (point 3, Proposition 2.1).

2.2.1 Classical ICE

Here, we assume that the distribution of a stationary PMC Z, which is defined by a density $p(z_1, z_2)$, depends on a parameter $\theta \in \Theta$. The problem is to estimate θ from a sample $y = (y_1, \dots, y_n)$. The classical Iterative Conditional Estimation (ICE) method is based on the following assumptions :

(i) there exists an estimator of θ from the complete data: $\hat{\theta} = \hat{\theta}(z) = \hat{\theta}((x_1, y_1), \dots, (x_n, y_n))$;
(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 is an iterative method which runs as follows :
1. Initialise $\theta = \theta^0$;
2. for $q \in N$,
-put $\theta^{q+1} = E_{\theta^q}[\hat{\theta}(Z)|Y = y]$ if the conditional expectation is computable,
- if not, simulate l realisations $x_1, \dots, x_l$ of X according to its distribution conditional to $Y = y$ and based on θ^q and put $\theta^{q+1} = \dfrac{\hat{\theta}(x_1, y) + \dots + \hat{\theta}(x_l, y)}{l}$.

As Z is stationary, its distribution is defined by $p_\theta(z_i, z_{i+1})$ which also is, according to the Proposition (2.1), the distribution of $U_i = (Z_i, Z_{i+1})$. Finding an estimator of θ from the complete data is, in general, not a serious problem (if it were, it would be pointless to seak an estimator of θ from the incomplete data, Whether by ICE or by any other method).
Let us specify how ICE runs in the three models of Examples 2.2 and 2.3. In the case of the Example 2.2 the parameters θ are the k^2 parameters $c_{ij} = p(\omega_i, \omega_j)$, $1 \le i, j \le k$, k means $(m_1, \dots, m_k)$ and k variances $(\sigma_1^2, \dots, \sigma_k^2)$ of the k Gaussian distributions $p(y_1|\omega_1)$, ..., $p(y_1|\omega_k)$. One possible $\hat{\theta} = \hat{\theta}(z)$ is then:

$$\hat{c}_{ij}(z) = \hat{c}_{ij}(x) = \frac{1}{n} \sum_{r=0}^{n} 1_{[(x_{2r-1}, x_{2r}) = (\omega_i, \omega_j)]} \quad (2.3)$$

$$\hat{m}_i(z) = \frac{1}{2n} \sum_{r=1}^{2n} 1_{[x_r = \omega_i]} y_r \quad (2.4)$$

$$\hat{\sigma}_i^2(z) = \frac{1}{2n} \sum_{r=1}^{2n} 1_{[x_r = \omega_i]} (y_r - \hat{m}_i(z))^2 \quad (2.5)$$

which relative simply frequencies and empirical means and variances.
Concerning the reestimation of $\hat{c}_{ij}$ given by (2.3), the conditional expectation is computable. One obtains

$$c_{ij}^{q+1} = E_{\theta^q}[\hat{c}_{ij}(X)|Y = y] = \frac{1}{n} \sum_{r=0}^{n} p(x_{2r-1}, x_{2r} = \omega_i, \omega_j|y)$$

which can be computed in a manner analogous to that of the *a posteriori* transition matrices in the proof of the Proposition 2.1). Concerning the reestimation of means and variances, the conditional expectation of (2.4) and (2.5) is not computable and one must resort to simulations of realisations of X according to its distribution conditional to $Y = y$ and based on θ^q.
When the first possibility of the example 2.3 is concerned (classical hidden Markov chain with correlated noise), we have to add to $(2k-1)/2$ parameters (c_{ij}), $1 \le i, j \le k$ (recall that $c_{ij} = c_{ji}$), k means and k variances, and $(2k-1)/2$ correlations of the $(2k-1)/2$ Gaussian

distributions $p(y_1, y_2 | \omega_i, \omega_j)$. (2.4) and (2.5) are then replaced by :

$$\hat{m}_{ij}(z) = \frac{1}{2n} \sum_{r=1}^{n} 1_{[(x_{2r-1}, x_{2r}) = (\omega_i, \omega_j)]} \begin{pmatrix} y_{2r-1} \\ y_{2r} \end{pmatrix} \tag{2.6}$$

$$\hat{\Sigma}_{ij}(z) = \tag{2.7}$$

$$= \frac{1}{2n} \sum_{r=1}^{2n} 1_{[(x_{2r-1}, x_{2r}) = (\omega_i, \omega_j)]} \left(\begin{pmatrix} y_{2r-1} \\ y_{2r} \end{pmatrix} - \hat{m}_{ij}(z) \right)^t \left(\begin{pmatrix} y_{2r-1} \\ y_{2r} \end{pmatrix} - \hat{m}_{ij}(z) \right)$$

so we have in this case $(2k-1)/2$ parameters (c_{ij}) and $(2k-1)/2$ variance-covariances matrices (Σ_{ij}) of the Gaussian distributions $p(y_1, y_2 | \omega_i, \omega_j)$. As above, the reestimations of (c_{ij}) is analytical and the reestimation of mean vectors (m_{ij}) and variance-covariances matrix (Σ_{ij}) are made through stochastic simulations.

2.2.2 Pairwise Markov Chain ICE (PMC-ICE)

The generalized ICE was first introduced in a hidden Markov discrete fields context, with application to unsupervised image segmentation [4], then generalised to any other discrete Hidden Markov Models [11], and finally extended to any discrete Hidden Markov Models with correlated sensors [18]. The ideas of [18] to deal with sensor dependencies is here applied to deal with spatial dependencies.

Let $Z = (X, Y)$ be a stationary PMC, with $X = \{\omega_1, \ldots, \omega_k\}$ and $Y = R$. Its distribution is defined by $p(z_1, z_2) = p(x_1, x_2) p(y_1, y_2 | x_1, x_2)$: thus the problem here is to estimate the $k(3k-1)/2$ probabilities $p(x_i, x_{i+1})$ and to find the $k(3k-1)/2$ probability densities $p(y_1, y_2 | x_1, x_2)$ on R^2. The aim of the generalized ICE is to seak the $k(3k-1)/2$ pdf's in as large a set as possible.

Let us consider the set Φ of M parametrized families of densities on R: $\Phi = \{F_1, \ldots, F_M\}$. For instance, F_1 may be Gaussian distributions, F_2 Gamma distributions, and so on, each family F_j being parametrized by a parameter β_j. We will then assume that for each (x_1, x_2), $p(y_1, y_2 | x_1, x_2)$ is the distribution of $\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a_{(x_1, x_2)} & 1 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$, where V_1 and V_2 are independent and the distribution of each of them belongs to one of the families in Φ.

Finally, we have to determinate the probabilities $p(x_i, x_{i+1})$ and, for each (x_1, x_2):

(i) estimate a;

(ii) determine i, j for which $P_{V_1} \in F_i$ and $P_{V_2} \in F_j$;

(ii) estimate the parameters β_i and β_j.

As mentioned above, the parameter estimation method we propose below is inspired from the "ICE-COR" method proposed in [18]. The class process studied in [18] was a Markov field and the random variables (Y_i) were assumed independent conditionally on X. However, there were two sensors: each Y_i took its values in R^2. For k classes the problem was to estimate a mixture of k distributions on R^2. Assuming that these distributions are neither independent nor Gaussian, we have a problem which is close enough, from the mathematical point of view, to the problem of estimating the distributions $p(y_1, y_2 | x_1, x_2)$ considered here.

So, let $Z = (X, Y)$ be a stationary PMC defined by $p_\theta(z_1, z_2)$, with $\theta \in \Theta_1 \cup \ldots \cup \Theta_M$, and let $\hat{\theta}_1 = \hat{\theta}_1(z)$, ..., $\hat{\theta}_M = \hat{\theta}_M(z)$ be M estimators, each $\hat{\theta}_l$ taking its values in Θ_l.

We assume having at our disposal a decision rule D such that for each z and $(\theta_1, \ldots, \theta_M) \in \Theta_1 \times \ldots \times \Theta_M$, $D(z) \in \{\theta_1, \ldots, \theta_M\}$.

The parameter estimation method we propose, which will be called PMC-ICE, is the following iterative algorithm:

1. Initialise $\theta = \theta^0$;
2. for $q \in N$,
-put $\theta_i^{q+1} = E_{\theta^q}[\hat{\theta}_i(Z) | Y = y]$ for $1 \le i \le M$ for which the conditional expectation is computable;
- for $1 \le i \le M$ for which the conditional expectation is not computable, simulate l realisations $x_1, \ldots, x_l$ of X according to its distribution conditional to $Y = y$ and based on θ^q and put $\theta_i^{q+1} = \frac{\hat{\theta}_i(x_1, y) + \ldots + \hat{\theta}_i(x_l, y)}{l}$;
-put $z_q = (x_1, y)$ and defined the next value of the parameter as $\theta^{q+1} = D(z_q) \in \{\theta_1^q, \ldots, \theta_M^q\}$.

This method of estimating θ can be of interest when we have M possible models for $p(z_1, z_2)$ (which are parametrized by Θ_1, ..., Θ_M, respectively) and we do not know in what case we are.

2.3 Numerous sensors

There is no theoretical limitation to consider several sensors. So, for m sensors we would have $X = \{\omega_1, \ldots, \omega_k\}$ and $Y = R^m$. For k classes we would then have to estimate a mixture of k^2 distribution on R^{2m}. So, there would be k^2 random vectors V, each having $2m$ independent components, and k^2 triangular matrices of size $2m \times 2m$. We can see how the estimation method proposed here extends the method proposed in [18], where the multivariate variables $Y_1 = (Y_1^1, \ldots, Y_1^m)$, ..., $Y_n = (Y_n^1, \ldots, Y_n^m)$ are independent conditionally on X. Here they are dependent conditionally on X and their distribution is given by the distribution of (Y_i, Y_{i+1}) conditional to (X_i, X_{i+1}). So, we have a more general model and recover again the model proposed in [18] when (Y_i, Y_{i+1}) are independent conditionally on (X_i, X_{i+1}).

3. 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 the idea having lead to the Pairwise Markov Random Field (PMRF) model recently proposed in [20]; however, there exist some significant differences between the two models.

We have discussed the differences and some possible advantages of the new model with respect to the classical Hidden Markov Chain (HMC) model. It appeared that in some situations it is possible to use the classical Bayesian restorations when using PMC, and it is not when using HMC.

The second point was to propose a method of PMC parameter estimation from the data Y alone. We described a fairly general method, valid even with non Gaussian or independent noise. The method was described in the one sensor case for clarity, but there is no theoretical limitation to consider multiple sensors.

Finally, we can say that the new PMC model and the corresponding parameter estimation method allows one to fuse, in an unsupervised manner, m pieces of information provided by m sensors and use the fused information in order to restore the observed noisy version of the process of interest.

As perspectives, let us mention the possibility of the use of HMC models in multiresolution images segmentation problem [8], which could thus be generalised to some PMC models. More generally, HMRF and HMC are particular cases of Hidden Markov models on networks [3, 25]. So, different generalisations of PMC proposed here - and PMRF proposed in [20] - to Pairwise Markov Processes on Networks could undoubtedly be considered and be possibly be of interest in some situations.

Acknowledgment. I thank Alain Hillion, Directeur Scientifique de l'École Nationale Supérieure des Télécommunications de Bretagne, for numerous discussions which greatly helped the writing of this paper.

References

[1] L.E. Baum, T. Petrie, G. Soules, N. Weiss. *A maximization technique occuring in the statstical analysis of probabilistic functions of Markov chains*, Ann, Math. Statistic., 41, pp. 164-171, 1970.

[2] J. Besag, *On the statistical analysis of dirty pictures*, Journal of the Royal Statistical Society, Series B, 48, pp. 259-302, 1986.

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

[4] Y. Delignon, A. Marzouki, and W. Pieczynski, *Estimation of Generalized Mixture and Its Application in Image Segmentation*, IEEE Transactions on Image Processing, Vol. 6, No. 10, pp. 1364-1375, 1997.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Series B, 39, pp. 1-38, 1977.

[6] P.A. Devijver, *Hidden Markov mesh random field models in image analysis*, Advances in Applied Statistics, Statistics and Images: 1, Carfax Publishing Company, pp. 187-227, 1993.

[7] A. El-Jacoubi, M. Gilloux, R. Sabourin, C. Y. Suen, *An HMM-Based Approach for Off-Line Unconstrained Handwritten Word Modeling and Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 8, pp. 752-760, 1999.

[8] L. Fouque, A. Appriou, and W. Pieczynski, *Multiresolution Hidden Markov Chain Model and Unsupervised Image Segmentation*, Proceedings of IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI'2000), pp. 121-125, 2-4 April 2000, Austin, Texas, USA.

[9] G.D. Fornay, *The Viterbi algorithm*, Proceedings of the IEEE, Vol. 61, No. 3, pp. 268-277, 1973.

[10] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*, IEEE Transactions on PAMI, Vol. 6, pp. 721-741, 1984.

[11] N. Giordana and W. Pieczynski, *Estimation of Generalized Multisensor Hidden Markov Chains and Unsupervised Image Segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 5, pp. 465-475, 1997.

[12] G. K. Kaleh and R. Vallet, J*oint parameter estimation and symbol detection for linear or nonlinear uknown channels,* IEEE Transactions on Communications, Vol. 42, No. 7, pp. 2406-2413, 1994.

[13] S. Lakshmanan, H. Derin, *Simultaneous parameter estimation and segmentation of Gibbs random fields*, IEEE Transactions on PAMI, Vol. 11, pp. 799-813, 1989.

[14] A. Maffet and C. Wackerman, *The modified Beta density function as a model for synthetic aperture radar clutter statistics*, IEEE Transactions on GRS, Vol. 29, No. 2, 1991.

[15] J. Marroquin, S. Mitter, T. Poggio, *Probabilistic solution of ill-posed problems in computational vision*, Journal of the American Statistical Association, 82, pp. 76-89, 1987.

[16] S. Medasani, R. Krishnapuram, *A Comparison of Gaussian and Pearson Mixture Modelling for Pattern Recognition and Computer Vision Applications*, Pattern Recognition Letters, 20, pp. 305-313, 1999.

[17] A. Peng, W. Pieczynski, *Adaptive Mixture Estimation and Unsupervised Local Bayesian Image Segmentation*, Graphical Models and Image Processing, Vol. 57, No. 5, pp. 389-399, 1995.

[18] W. Pieczynski, J. Bouvrais, and C. Michel, *Estimation of Generalized Mixture in the Case of Correlated Sensors*, IEEE Transactions on Image Processing, Vol. 09, No. 2, pp. 308-312, 2000.

[19] W. Pieczynski, *Statistical image segmentation, Machine Graphics and Vision*, Vol. 1, No. 1/2, pp. 261-268, 1992.

[20] W. Pieczynski and A.-N. Tebbache, *Paiwise Markov random fields and their application in textured images segmentation*, Proceedings of IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI'2000), pp. 106-110, 2-4 April 2000, Austin, Texas, USA.

[21] W. Qian, D.M. Titterington, *On the use of Gibbs Markov chain models in the analysis of images based on second-order pairwise interactive distributions*, Journal of Applied Statistics, Vol. 16, No. 2, pp. 267-282, 1989.

[22] L.R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of IEEE, Vol. 77, No. 2, pp. 257-286, 1989.

[23] C. Raphael, *Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21 No. 4, pp. 360-370, 1999.

[24] F. Salzenstein et W. Pieczynski, *Sur le Choix de Méthode de Segmentation Statistique d'Images*, Traitement du Signal, Vol. 15, No. 2, pp. 119-128, 1998.

[25] R. Serfozo, *Introduction to Stochastic Networks*, Springer-Verlag, New York, 1999.

[26] A. D. Wilson, A. F. Bobick, *Parametric Hidden Markov Models for Gesture Recognition*, IEEE Transactions on Image Processing, Vol. 8 No. 9, pp. 884-900, 1999.