Unsupervised segmentation by hidden Markov chain with bi-dimensional observed process

Abdelali Joumad and Abdelaziz Nasroallah

Abstract—In an unsupervised segmentation context, we propose a bi-dimensional hidden Markov chain model (BHMC) that we adapt to the image segmentation problem. The bi-dimensional observed process \( Y = (Y^1, Y^2) \) is such that \( Y^1 \) represents the noisy image and \( Y^2 \) represents a noisy supplementary information on the image, for example a noisy proportion of pixels of the same type in a neighborhood of the current pixel. The proposed model can be seen as a competitive alternative to the Hilbert-Peano scan. We propose a bayesian algorithm to estimate parameters of the considered model. The performance of this algorithm is globally favorable, compared to the bi-dimensional EM algorithm through numerical and visual data.

Keywords—Image segmentation, Hidden Markov chain with a bi-dimensional observed process, Peano-Hilbert scan, Bayesian approach, MCMC methods, Bi-dimensional EM algorithm.

I. INTRODUCTION

HIDDEN Markov models (HMM) are frequently used to study one unobserved variable and uni-dimensional observed process such as hidden Markov fields [13], [28], [8], [5], [30], [40], [10], [27], [35], [17], hidden Markov chain [7], [13], [5], [22], [30], [12], [36], pairwise hidden Markov model [32] or more unobserved variable and uni-dimensional observed process such as triplet hidden Markov model [11]. The multivariate observed process in HMM is used to study supervised multi-component images segmentation. Such vectorial image can be obtained for example, from different channels, from several sensors or from image taken at various moments. There are several interesting publications, like Spherically Invariant Random Vector models (SIRV) and copula [15], [21], [24], [33], [34], that treat the multi-dimensional case [22], [24], [23], [31]. One of the main problems in image modeling is the adequacy between temporal and spatial context of pixels during the image modeling. This difficulty is usually solved by Markov fields, or by Hilbert-Peano scan in Markov chain case [7], [23]. We propose a hidden Markov chain with a bi-dimensional observed process as another alternative to study this problem. The cited works above concerned with multi-component images use an observed process \( Y = (Y^1, \ldots, Y^M) \) where \( \Omega = \{\omega_1, \ldots, \omega_K\} \) is the state space of Markov process and \( K \geq 2 \) is a number of classes. The process \( X \) is completely characterized by transition parameters \( c_{ij} = P(X_t = \omega_i, X_{t+1} = \omega_j) \) for \( i, j = 1, \ldots, K \); the initial law \( \nu = (\nu_i, 1 \leq i \leq K) \) and the transition matrix \( A = (a_{ij})_{1 \leq i, j \leq K} \) are such that

\[
\nu_i = P(X_{t_1} = \omega_i) = \sum_{1 \leq j \leq K} c_{ij} \quad \text{and} \quad a_{ij} = \frac{c_{ij}}{1 \leq j \leq K},
\]

where \( t_1 \) indicates the initial time of Markov chain \( X \) that is supposed homogenous. Furthermore we suppose the following hypothesis:

\( (H_1) \) : \( \forall (t, t') \in \mathbb{N}^2 \) such that \( t \neq t' \), the conditional processes \( Y_{t'|X} \) and \( Y_{t'|X} \) are independent.

\( (H_2) \) : \( \forall t \in \mathbb{N} \), \( Y_{t}|X \) and \( Y_{t}|X \) have the same gaussian distribution.

So \( \forall t \in \mathbb{N} \), \( Y_{t}|X_t = \omega_i \sim \mathcal{N}(\mu_i, \Sigma_i) \), where the density function is given by

\[
f_i(y_t) = \frac{1}{2\pi|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(y_t - \mu_i)\Sigma_i^{-1}(y_t - \mu_i)\right).
\]
with $\mu_i = (\mu_{1i}, \mu_{2i})'$ and $\Sigma_i = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{pmatrix}$, $i = 1, \ldots, K$, are respectively the mean vector and variance-covariance matrix of bi-normal distribution, and $'*$ is the transpose operator. The process $(X, Y)$ is called bi-dimensional hidden Markov chain (BHMC), where $X$ is the unobserved process and $Y = (Y^1, Y^2)$ is the observed one. For a realization of $(X, Y)$

$$(x, y) = ((\omega_1, y_1), \ldots, (\omega_n, y_n)),$$

the likelihood is given by

$$L(x, y; \Theta) = \nu_1 \prod_{m=2}^n \alpha_{m-1|m} \prod_{m=1}^n f_{im}(y_m),$$

where $\Theta = (\nu, A, (\mu_1, \Sigma_1), \ldots, (\mu_K, \Sigma_K))$. The size of $\Theta$ is $K^2 + 7K$ (i.e. $K$ for $\nu$, $K^2$ for $A$, $2K$ for $(\mu_i, \Sigma_i)$, $i = 1, \ldots, K$ and $4K$ for $(\Sigma_i)$). Using the fact that each row of $A$ and $\nu$ are probability distributions and the symmetry of $\Sigma$, this size can be reduced to $K^2 + 5K - 1$. We suppose that block components $\nu, A, (\mu_1, \Sigma_1), i = 1, \ldots, K$ of $\Theta$ are independent.

The classification problem is thus the estimation of the unobserved process realization $X = x$ from the observation $Y = y$. The Bayesian Maximum A Posteriori (MAP) and Maximum Posterior Mode (MPM) can be easily computed for the BHMC model. In this paper we will use the MPM method based on minimizing the mean ratio of wrongly classified points. Formally, for each $t$, the state $\omega_t$ that maximizes the a posteriori probability is such that

$$P(X_t = \omega_j \mid Y = y) = \max_{1 \leq i \leq K} P(X_t = \omega_i \mid Y = y)$$

The computation of the posterior marginal distribution is feasible thanks to the “Forward-Backward” method. Finally, the problem is: given the bi-dimensional observed process $Y$, we have to estimate the global parameter $\Theta$.

### III. BEM Algorithm

Generally, the maximum likelihood estimator has good statistical properties. The Expectation-Maximization (EM) algorithm [2], [26], [37] is one of the most popular method to approximate parameter $\Theta$ maximizing the likelihood $L(x, y; \Theta)$. Given an initial guess $\Theta^{(0)}$, the EM algorithm consists on generating iteratively a sequence $(\Theta^{(q)})_{q \in \mathbb{N}}$ that globally converges to the theorem 10.5.2 in [25]. The two principal steps of EM are:

**E-step:** Determine $Q(\Theta, \Theta^{(q)}) = E_{\Theta^{(q)}}(\ln L(x, y; \Theta))|y$.

**M-step:** Choose $\Theta^{(q+1)}$ to be the value that maximizes $Q(\Theta, \Theta^{(q)})$.

To adapt the EM algorithm to the hidden Markov chain with a bi-dimensional observed process, we need the following notations and definitions. For $t \in \mathbb{N}$ and $\omega_i, \omega_j \in \Omega$, we define:

$$\psi_t(i, j) := P(X_t = \omega_i, X_{t+1} = \omega_j \mid Y = y)$$

and

$$\chi_t(i) := P(X_t = \omega_i \mid Y = y).$$

These probabilities can be expressed according to the forward-backward probabilities

$$\alpha_t(i) = P(X_t = \omega_i \mid Y_1 = y_1, \ldots, Y_t = y_t)$$

and

$$\beta_t(i) = \frac{P(Y_{t+1}, \ldots, Y_N = y_{t+1}, \ldots, y_N \mid X_t = \omega_i)}{\sum_{j \in \Omega} P(Y_{t+1}, \ldots, Y_N = y_{t+1}, \ldots, y_N \mid X_t = \omega_j) \beta_{t+1}(j)}.$$
A. Full conditional distribution for $A$

Using (1), for $i,j = 1, \ldots, K$, the estimation of $a_{ij}$ can be deduced from those of $c_{ij}$. So it’s being sufficient to determine the full conditional distribution of $c_{ij}$.

For $u,v = 1, \ldots, K$,

$$\pi(c_{uv} | \cdot) \propto \pi(c_{uv}) \pi(Y \mid X = x, \nu, A, \mu, \Sigma) \propto \pi(c_{uv}) \pi(X = x | \nu, A, \mu, \Sigma),$$

where $\mu = (\mu_i; 1 \leq i \leq K)$, $\Sigma = (\Sigma_{ij}; 1 \leq i \leq K)$ and the symbol $\propto$ indicates proportionality up to a constant that does not depend on the considered variable.

On the other hand, including the fact that $X$ is an homogenous Markov chain and $x = (x_{ik})_{1 \leq i \leq r, 1 \leq k \leq r_i}$ is a realization of $r$ independent trajectories of $X$, we have

$$P(X = x | \nu, A) = \prod_{i=1}^{r} P(X_{it1} = x_{i1} | \nu) \prod_{k=2}^{r_i} P(X_{itk} = x_{ik} | X_{itk-1} = x_{i(k-1), \nu, A})$$

So

$$\pi(c_{uv} | \cdot) \propto \pi(c_{uv}) \prod_{i=1}^{r} \nu_{x_{i1}}^{2-r_i} \prod_{k=2}^{r_i} P(X_{itk} = x_{ik} | X_{itk-1} = x_{i(k-1), \nu, A}),$$

To select the prior distribution $\pi(c_{uv})$, Gamma$(\alpha_{uv}, 1)$ is a natural choice since it leads to a Dirichlet distribution supposed for each row of the transition matrix $A$. The generation of the univariate full conditional distribution for $c_{uv}$ can be obtained by a Metropolis accept-reject algorithm.

**Remark 1.** Simulation from a Dirichlet distribution $D_K(\delta_1, \ldots, \delta_K)$ can be drawn thanks to the following result: if $\xi_1, \ldots, \xi_K$ are independent with $\xi_i$ having a Gamma$(\delta_i, 1)$ distribution, then

$$\left( \frac{\xi_1}{\sum_{i=1}^{K} \xi_i}, \frac{\xi_2}{\sum_{i=1}^{K} \xi_i}, \ldots, \frac{\xi_K}{\sum_{i=1}^{K} \xi_i} \right) \sim D_K(\delta_1, \ldots, \delta_K)$$

B. Full conditional distribution for $\nu$

The full conditional initial distribution of Markov chain is given by:

$$\pi(\nu | \cdot) \propto \pi(\nu) \prod_{i=1}^{r} \nu_{x_{i1}}^{2-r_i},$$

which can be written as

$$\pi(\nu | \cdot) \propto \pi(\nu) \prod_{k=1}^{K} r_k^{r_k(1)},$$

where $r_k(1) := \sum_{i=1}^{r_i} \mathbb{1}_{\{x_{ik} = \omega_j\}}$ is the number of visits to state $\omega_j$ at initial time ($\mathbb{1}$ is the indicator function of set $E$). A conjugate prior for $\nu = (\nu_1, \ldots, \nu_K)$ is a Dirichlet distribution $D_K(\delta_1, \ldots, \delta_K)$ with known hyper-parameters $\delta_k > 0, k = 1, \ldots, K$ (i.e. $\pi(\nu) \propto \prod_{k=1}^{K} \nu_k^{\delta_k-1}$).

This choice is justified by the fact that $\nu_1, \ldots, \nu_K$ are parameters of a multinomial distribution. So we obtain

$$\pi(\nu | \cdot) \propto \prod_{k=1}^{K} \nu_k^{r_k(1)+\delta_k-1},$$

which leads to

$$\pi(\nu | \cdot) \sim D_K(\delta_1 + r_1(1), \ldots, \delta_K + r_K(1)).$$

C. Full conditional distribution for $\mu_j$ and $\Sigma_j$

We suppose that $\Sigma_j$, for $j = 1, \ldots, K$, are independent and $\mu_j$ depends only on $\Sigma_j$, $j = 1, \ldots, K$, the corresponding prior distributions are chosen to be

$$\mu_j \mid \Sigma_j \sim N_2(\tau_j, \Sigma_j) \quad \text{and} \quad \Sigma_j \sim Inv - Wishart(\beta_j, W_j^{-1}),$$

where the hyper-parameters $\tau_j, \beta_j$ and $W_j^{-1}$ are supposed known.

Let’s consider the following definitions and notations: $N_i^j := \{k = 1, \ldots, r_i, x_{ik} = \omega_j\}$ is the set of times of visits to state $\omega_j$ in trajectory $i$, $N_j := \sum_{i=1}^{r} | N_i^j |$, $\overline{y}_j = \frac{1}{N_j} \sum_{i=1}^{r} \sum_{k \in N_i^j} y_{ik}$ and $S_j = \sum_{i=1}^{r} \sum_{k \in N_i^j} (y_{ik} - \overline{y}_j)(y_{ik} - \overline{y}_j)^\prime$, where $| N_i^j |$ is the cardinality of the finite set $N_i^j$ and $y = (y_{ik})_{1 \leq i \leq r, 1 \leq k \leq r_i}$ is the given realization of $Y$.

So, for $j = 1, \ldots, K$,

$$\pi(\Sigma_j, \mu_j | \cdot) \propto \pi(\Sigma_j, \mu_j) \pi(Y | X = x, \Sigma, \mu) \propto \pi(\Sigma_j) \pi(\mu_j | \Sigma_j) \pi(Y | X = x, \Sigma, \mu).$$

On the other hand, by hypothesis $(H_1)$ and $(H_2)$ in section II, we have

$$\pi(Y | X = x, \Sigma, \mu) \propto \prod_{i=1}^{r} \prod_{k=1}^{r_i} \pi(Y_{itk} | X_{itk} = x_{ik}, \Sigma_{x_{ik}}, \mu_{x_{ik}})$$

$$\propto \prod_{i=1}^{r} \prod_{k=1}^{r_i} | \Sigma_{x_{ik}} |^{-\frac{1}{2}} e^{-\frac{1}{2} (Y_{itk} - \mu_{x_{ik}})' \Sigma_{x_{ik}}^{-1} (Y_{itk} - \mu_{x_{ik}})}.$$

Therefore, the full conditional distribution of $(\Sigma_j, \mu_j)$ is
given by
\[
\pi(\Sigma_j, \mu_j) \propto \pi(\Sigma_j) \pi(\mu_j | \Sigma_j) \prod_{i=1}^{r} | \Sigma_j |^{-\frac{N}{2}} e^{-\frac{1}{2} (y_{it-k} - \mu_j)^T \Sigma_j^{-1} (y_{it-k} - \mu_j)}
\]
\[
\propto \pi(\Sigma_j) \frac{\pi(\mu_j | \Sigma_j)}{|\Sigma_j|^{\frac{N}{2}}} e^{-\frac{1}{2} tr(\Sigma_j^{-1} (S_j + N_j (\bar{y}_j - \mu_j)^T (\bar{y}_j - \mu_j)))}
\]
\[
= \left( \pi(\Sigma_j) \right)^{\frac{1}{2}} \left( \pi(\mu_j | \Sigma_j) \right)^{\frac{1}{2}} e^{-\frac{1}{2} [tr(\Sigma_j^{-1} W_j^{-1}) + tr((\Sigma_j^{-1} \bar{y}_j)(\Sigma_j^{-1} \bar{y}_j)^T)]}
\]
\[
\times e^{-\frac{1}{2} [tr((\Sigma_j^{-1} \bar{y}_j)(\Sigma_j^{-1} \bar{y}_j)^T)]}
\]

Thereby
\[
\pi(\mu_j | \Sigma_j) \sim N_2 \left( \tau_j + N_j \bar{y}_j, \frac{\Sigma_j}{N_j + 1} \right)
\]
and
\[
\pi(\Sigma_j) \sim Inv \left( \mu_j | \rho_j \right)
\]

with
\[
W(S_j, W_j) = W_j^{-1} + S_j + \frac{N_j}{N_j + 1} (\bar{y}_j - \tau_j)(\bar{y}_j - \tau_j)^T.
\]

Remark 2. As argued in [9], [25], often there is no compelling reason to choose the previous priors, except for their simplicity, but the restrictive aspect of conjugate priors can be attenuated by using hyper-priors on the hyper-parameters. Those hyper-priors can be chosen amongst so-called non-informative (or vague) priors to attenuate the impact on the resulting inference.

Now, the joint prior distribution for \( \Theta \) can be written as:
\[
\pi(\Theta) = \pi(\nu_1, \ldots, \nu_K, A, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)
\]
\[
= \pi(\nu_1, \ldots, \nu_K) \pi(A) \pi(\Sigma_1, \ldots, \Sigma_K)
\]
\[
= \prod_{j=1}^{K} \frac{1}{\Gamma(\nu_j)} \pi(\mu_j | \Sigma_j)
\]
\[
\propto \prod_{j=1}^{K} \frac{1}{\Gamma(\nu_j)} \pi(\mu_j | \Sigma_j)
\]
\[
\times \prod_{j=1}^{K} \pi(\Sigma_j) \prod_{j=1}^{K} (\Sigma_j)^{-\frac{1}{2}} e^{-\frac{1}{2} (\mu_j - \tau_j)(\Sigma_j^{-1})(\mu_j - \tau_j)}
\]
\[
\times \prod_{j=1}^{K} (\Sigma_j)^{-\frac{1}{2}} e^{-\frac{1}{2} [tr(\Sigma_j^{-1} W_j^{-1})]}
\]

This joint prior distribution is obtained using independence of block components of \( \Theta \) supposed in section II.

D. A bayesian algorithm for BHMC

We propose a bayesian algorithm for bi-dimensional hidden Markov chain model, principally based on the classical Gibbs sampler and Metropolis accept-reject algorithms. For given hyper-parameters \( \alpha_j, \tau_j, \delta_j, \beta_j, \) and \( W_j^{-1} ; i = 1, \ldots, K, \) the main steps of BBHMC are the following:

The BBHMC Algorithm

E0. Select starting values (iteration \( q = 0 \)),
\[
\Theta^{(0)} = (\nu_1^{(0)}, \mu_1^{(0)}, \Sigma_1^{(0)}, \ldots, \nu_K^{(0)}, \mu_K^{(0)}, \Sigma_K^{(0)}, i, j, 1, \ldots, K)
\]
and compute \( c_{ij}^{(0)} \) using (1).

E1. Iteration \((q + 1)\)

E1.1 Generate a sample \((\bar{x}^{(q + 1)}), i = 1, \ldots, r, k = 1, \ldots, r_1\)

E1.2 Generate a transition matrix \( A^{(q + 1)} = \{ a_{ij}^{(q + 1)} \}_{1 \leq i, j \leq K} \)

E1.2.1 Generate a candidate \( b_{ij} \sim N(c_{ij}^{(q)}, 1) \) and \( u \sim U(0, 1) \)

such that \( b_{ij} \) and \( u \) are independent.

E1.2.2 Compute \( \prod(b_{ij}) \) and \( \prod(c_{ij}^{(q)}) \) by

\[
\prod(b_{ij}) = \prod_{i=1}^{r_1} \prod_{k=1}^{r_1} [q_{ij}^{(q)}] \prod_{i=1}^{r_1} [\tilde{x}_{i(i-1)}(i)]
\]
\[
\prod(c_{ij}^{(q)}) = \prod_{i=1}^{r_1} [q_{ij}^{(q)}] \prod_{i=1}^{r_1} [\tilde{x}_{i(i-1)}(i)]
\]

E1.2.3 Compute \( \rho_{ij}^{(q + 1)} = \min \left( \frac{q_{ij}^{(q)}}{q_{ij}^{(q + 1)}}, \frac{q_{ij}^{(0)}}{q_{ij}^{(q + 1)}} \right) \) where \( \gamma \) is the density function of \( Gamma(\alpha_{ij}, 1) \)

E1.2.4 Compute \( c = c_{ij}^{(q)} \) if \( c = c_{ij}^{(q)} \), we reject \( b_{ij} \) and return to E1.2.1
else we accept \( b_{ij} : c_{ij}^{(q + 1)} = c \)

E1.2.5 Compute \( a_{ij}^{(q + 1)} \) (by formula in (1))

E2. Compute \( N_j, S_j, \bar{y}_j \) and \( W(S_j, W_j) \) for \( j = 1, \ldots, K \)

E3. For \( j = 1, \ldots, K \),

E3.1 Generate

\[
\pi(\mu_j^{(q + 1)} | \cdot) \sim N_2 \left( \tau_j + N_j \bar{y}_j, \frac{\Sigma_j^{(q)}}{N_j + 1} \right)
\]

E3.2 Generate

\[
\pi(\Sigma_j^{(q + 1)} | \cdot) \sim Inverse - Wishart(\beta_j + N_j, W(S_j, W_j))
\]

E3.3 Generate

\[
\pi(\nu^{(q + 1)} | \cdot) \sim D_K \left( \delta_1 + r_1(1), \ldots, \delta_K + r_K(1) \right)
\]

V. Numerical simulation

We interest to the performance of BEM and BBHMC algorithms by simulating different noisy Markov chains, and we compare the estimations obtained from the noisy data to those computed from the complete data. These last will be called
Empirical Values and will be denoted (EV). The computation of the empirical values $c_{ij}^*, \mu_j^*$ and $\Sigma_j^*$ for $i, j = 1, \ldots, K$ makes itself by the following formulas:

$$c_{ij} = \frac{1}{r(r-1)} \sum_{l=1}^{r} \sum_{k=1}^{r-1} \prod_{h \in \{X_k, X_{l(k+1)}\}}(w_i, w_j),$$

$$\mu_j^* = \frac{1}{N_j} \sum_{k \in N_j} \sum_{i \in X_k} y_{itk},$$

$$\Sigma_j^* = \frac{1}{N_j} \sum_{k \in N_j} \sum_{i \in X_k} (y_{itk} - \mu_j^*)(y_{itk} - \mu_j^*)'.$$

In the following, we present the parameter initializations and the data scenarios, that has been used by many authors [7], [14], [29], after adapting them to the bi-dimensional case.

A. Parameter initializations

For $i, j = 1, \ldots, K$ and $d, d' = 1, 2$

- Markov chain parameters:
  $$\nu_i = \frac{1}{K}, \quad a_{ii} = 0.5 \quad \text{and} \quad a_{ij} = \frac{1}{2(K-1)} \quad \text{for} \ i \neq j$$

- Noisy parameters:
  Mean vector:
  If $K = 2p$, for $1 \leq k \leq p$
  $$\mu_{k+1}^d = m^d - 0.5(p-k)\sqrt{\sigma_{dd}}$$
  $$\mu_{K-k}^d = m^d + 0.5(p-k)\sqrt{\sigma_{dd}}$$
  If $K = 2p + 1$, for $1 \leq k \leq K$
  $$\mu_{k+1}^d = m^d - 0.5(0.5K - k)\sqrt{\sigma_{dd}}.$$  
  Variance-covariance matrix:
  $$\sigma_{dd}^a = \sigma_{dd}, \quad \sigma_{dd'}^a = c_1 \sigma_{dd} \quad \text{for} \ d \neq d'$$
  where $c_1 = 0.1$ for numerical simulation and $c_1 = 0.9$ for image segmentation. The mean vector $(m^1, m^2)$ and the variance-covariance matrix $\left(\begin{array}{cc} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{array}\right)$ are computed using a generated trajectory of $X$.

- Hyper-parameters:
  For $1 \leq i, j \leq K$, $\alpha_{ij} = 2$, $\tau_i = (2, 2)'$, $\beta_i = 1$, $\delta_i = 1$, $1$, and $W_i^{-1} = \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right)$

B. Markov chain and noisy data scenarios

While considering a sample of size $n = 1000$ and a number of classes $K = 2$, there are two factors of wish we study the influence on the performance of BEM and BBHMC algorithms: the noise and the Markov chain.

- For the noise factor, we consider the following three scenarios:
  \begin{itemize}
    \item **(MD):** means discriminating noise $\mu_1 \neq \mu_2$ and $\Sigma_1 = \Sigma_2$
    \item **(VD):** variances discriminating noise $\mu_1 = \mu_2$ and $\Sigma_1 \neq \Sigma_2$
    \item **(MVD):** means and variances discriminating noise $\mu_1 \neq \mu_2$ and $\Sigma_1 \neq \Sigma_2$
  \end{itemize}

The noise level depends strongly on the following two parameters:

$$\Delta = (\mu_2^1 - \mu_1^1)^2 + (\mu_2^2 - \mu_1^2)^2 \quad \text{and} \quad \rho = \frac{\rho_2}{\rho_1},$$

where

$$\rho_2 = \frac{\sigma_{12}^2}{\sqrt{\sigma_{11}^2 \sigma_{22}^2}} \quad \text{and} \quad \rho_1 = \frac{\sigma_{12}^2}{\sqrt{\sigma_{11}^2 \sigma_{22}^2}}.$$

We fix the parameters values for each scenario as follows:

- **MD:** $\mu_1 = (1, 1)'$, $\mu_2 = (1, 2)'$ and $\Sigma_1 = \Sigma_2 = \left(\begin{array}{cc} 2 & 0.9 \\ 0.9 & 3 \end{array}\right)$.
  These values correspond to $(\Delta, \rho) = (1, 1)$.
- **VD:** $\mu_1 = \mu_2 = (1, 1)'$, $\Sigma_1 = \left(\begin{array}{cc} 0.25 & 0.2 \\ 0.2 & 4 \end{array}\right)$ and $\Sigma_2 = \left(\begin{array}{cc} 5 & 5.3 \\ 5.3 & 9 \end{array}\right)$.
  These values correspond to $(\Delta, \rho) = (0, 1.56)$.
- **MVD:** $\mu_1 = (1.1, 1.5)'$, $\mu_2 = (2.1, 2.5)'$, $\Sigma_1 = \left(\begin{array}{cc} 0.25 & 0.9 \\ 0.9 & 4 \end{array}\right)$ and $\Sigma_2 = \left(\begin{array}{cc} 5 & 3.5 \\ 3.5 & 9 \end{array}\right)$.
  These values correspond to $(\Delta, \rho) = (1.25, 0.30)$.

- For the Markov chain factor, we consider the following two scenarios:
  \begin{itemize}
    \item **C1:** $a_{11} = a_{22} = 0.98$
    \item **C2:** $a_{11} = a_{22} = 0.80$
  \end{itemize}

C. Implementation parameters

There are several parameters which arise when implementing MCMC algorithms [39]. The algorithm must be run for a heating period, called burn-in, which depends on initial values of parameters. The aim of this period is to reduce the effect of initial values on posterior inference. Visual inspection of plots of the algorithm output $(\Theta^{(q)}, q = 0, 1, \ldots, T)$ with $T$ is the global length of iterations, is the most obvious and commonly used method for determining the burn-in period. For the burn-in period length $T_0$ of BBHMC algorithm, we take $T_0 = 1000$ for numerical simulations and $T_0 = 7000$ for image segmentation. After this period, we start the collection of estimations for $T_1 = 1000$ supplementary iterations. So, the total number of iterations is $T = T_0 + T_1$ and the estimation $\hat{\Theta}$ of $\Theta$ is given by the MMSE estimator $(\text{i.e.} \quad \hat{\Theta} = \frac{1}{T - T_0 + 1} \sum_{q=T_0}^{T} \Theta^{(q)})$.

For BEM algorithm, we use an automatic stopping criteria based on the following distances (inspired to [7])

$$D^2(\mu^{(q)}) = \frac{1}{K} \sum_{k=1}^{K} c_k \left[ (\mu_k^{(1)})^2 - \frac{1}{q+1} \sum_{l=0}^{q} \mu_k^{(l)2} \right] + \left[ (\mu_k^{(2)})^2 - \frac{1}{q+1} \sum_{l=0}^{q} \mu_k^{(l)2} \right]$$

$$D^2(\sigma_{ij}^{(q)}) = \frac{1}{K} \sum_{k=1}^{K} c_k \left[ \sigma_{ij}^{(k)} - \frac{1}{q+1} \sum_{l=0}^{q} \sigma_{ij}^{(l)} \right] \quad ; \quad 1 \leq i, j \leq 2$$

The stopping of iterations is a function of quality estimation of parameters and stoutness of MPM method. The BEM
algorithm is stopped when $D^2(\mu(q)) < \varepsilon$ and $D^2(\sigma(q)) < \varepsilon$ for $i, j = 1, 2$. We take $(\varepsilon, c_k) = (10^{-4}, 1)$ for numerical simulations and $(\varepsilon, c_k) = (10^{-2}, 10^{-3})$ for image segmentation.

### D. Numerical results

This subsection deals with numerical estimation for parameters and misclassification rates obtained by BEM and BBHMC, compared with those obtained by EV. These algorithms are tested on combinations of C1 and C2 with MD, VD and MVD. The BEM and BBHMC algorithms use the same initializations. The obtained estimations are presented in Tables II to V. In Table I, we present the misclassification rates of pixels computed by MPM method.

From Table I, we remark that BBHMC has globally a small advantage with respect to BEM mainly for C2+MD and C2+MVD scenarios with which are counted as worst cases.

Concerning estimations of $\nu$ and $A$, resumed in Tables II and III, the BEM algorithm takes advantage with respect to BBHMC algorithm with a stability advantage of estimations for the last. It seems that this stability is due to the estimations form (i.e. estimators obtained by the law number large).

Regarding Tables IV and V, the estimations of $\mu$ and $\Sigma$ obtained by BEM and BBHMC are very close to those obtained by EV.

### TABLE I

**Error rates for each class (ER$_i$ and E$_{R_i}$) given by EV, BEM and BBHMC for different data scenarios, where ER$_i$ is the misclassification rate of class i. The mean CPU time for BEM and BBHMC is 7.5.**

<table>
<thead>
<tr>
<th>Data</th>
<th>Error rate(%)</th>
<th>EV</th>
<th>BEM</th>
<th>BBHMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1-MD</td>
<td>$E_{R_i}$</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>$E_{R_2}$</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
</tr>
<tr>
<td>C1-VD</td>
<td>$E_{R_i}$</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>$E_{R_2}$</td>
<td>1.8</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>C1-MVD</td>
<td>$E_{R_1}$</td>
<td>0.1</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>$E_{R_2}$</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>C2-MD</td>
<td>$E_{R_1}$</td>
<td>22.6</td>
<td>23.7</td>
<td>22.6</td>
</tr>
<tr>
<td></td>
<td>$E_{R_2}$</td>
<td>18.5</td>
<td>19.0</td>
<td>18.5</td>
</tr>
<tr>
<td>C2-VD</td>
<td>$E_{R_1}$</td>
<td>6.7</td>
<td>7.1</td>
<td>6.9</td>
</tr>
<tr>
<td></td>
<td>$E_{R_2}$</td>
<td>12.7</td>
<td>12.9</td>
<td>12.7</td>
</tr>
<tr>
<td>C2-MVD</td>
<td>$E_{R_1}$</td>
<td>4.8</td>
<td>4.2</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>$E_{R_2}$</td>
<td>12.9</td>
<td>16.0</td>
<td>12.9</td>
</tr>
</tbody>
</table>

### TABLE II

**Estimations of component $a_{ij}$ by EV, BEM and BBHMC for different data scenarios. The standard deviation is given between brackets for BBHMC estimations.**

<table>
<thead>
<tr>
<th>Data</th>
<th>EV</th>
<th>BEM</th>
<th>BBHMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1-MD</td>
<td>0.522</td>
<td>0.523</td>
<td>0.490 (0.410×10$^{-2}$)</td>
</tr>
<tr>
<td>C1-VD</td>
<td>0.522</td>
<td>0.523</td>
<td>0.490 (0.410×10$^{-2}$)</td>
</tr>
<tr>
<td>C1-MVD</td>
<td>0.522</td>
<td>0.519</td>
<td>0.490 (0.410×10$^{-2}$)</td>
</tr>
<tr>
<td>C2-MD</td>
<td>0.475</td>
<td>0.477</td>
<td>0.493 (0.021×10$^{-2}$)</td>
</tr>
<tr>
<td>C2-VD</td>
<td>0.475</td>
<td>0.482</td>
<td>0.493 (0.021×10$^{-2}$)</td>
</tr>
<tr>
<td>C2-MVD</td>
<td>0.475</td>
<td>0.501</td>
<td>0.493 (0.021×10$^{-2}$)</td>
</tr>
</tbody>
</table>

### TABLE III

**Estimations of $a_{11}$ and $a_{22}$ by EV, BEM and BBHMC for different data scenarios. The standard deviation is given between brackets for BBHMC estimations.**

<table>
<thead>
<tr>
<th>Data</th>
<th>EV</th>
<th>BEM</th>
<th>BBHMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1-MD</td>
<td>$a_{11}$</td>
<td>0.980</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td>$a_{22}$</td>
<td>0.979</td>
<td>0.978</td>
</tr>
<tr>
<td>C1-VD</td>
<td>$a_{11}$</td>
<td>0.980</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>$a_{22}$</td>
<td>0.979</td>
<td>0.978</td>
</tr>
<tr>
<td>C1-MVD</td>
<td>$a_{11}$</td>
<td>0.980</td>
<td>0.978</td>
</tr>
<tr>
<td></td>
<td>$a_{22}$</td>
<td>0.979</td>
<td>0.978</td>
</tr>
<tr>
<td>C2-MD</td>
<td>$a_{11}$</td>
<td>0.791</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td>$a_{22}$</td>
<td>0.809</td>
<td>0.808</td>
</tr>
<tr>
<td>C2-VD</td>
<td>$a_{11}$</td>
<td>0.791</td>
<td>0.800</td>
</tr>
<tr>
<td></td>
<td>$a_{22}$</td>
<td>0.809</td>
<td>0.807</td>
</tr>
</tbody>
</table>

### TABLE IV

**Estimations of the two component of $a_{12}$ and $a_{21}$ by EV, BEM and BBHMC for different data scenarios. The standard deviation is given between brackets for BBHMC estimations.**

<table>
<thead>
<tr>
<th>Data</th>
<th>EV</th>
<th>BEM</th>
<th>BBHMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1-MD</td>
<td>$\mu_1$</td>
<td>0.984</td>
<td>1.064</td>
</tr>
<tr>
<td></td>
<td>$\mu_2$</td>
<td>2.031</td>
<td>2.344</td>
</tr>
<tr>
<td>C1-VD</td>
<td>$\mu_1$</td>
<td>0.981</td>
<td>1.056</td>
</tr>
<tr>
<td></td>
<td>$\mu_2$</td>
<td>2.049</td>
<td>2.046</td>
</tr>
<tr>
<td>C1-MVD</td>
<td>$\mu_1$</td>
<td>0.981</td>
<td>1.056</td>
</tr>
<tr>
<td></td>
<td>$\mu_2$</td>
<td>2.018</td>
<td>2.044</td>
</tr>
<tr>
<td>C2-MD</td>
<td>$\mu_1$</td>
<td>0.982</td>
<td>1.040</td>
</tr>
<tr>
<td></td>
<td>$\mu_2$</td>
<td>1.992</td>
<td>1.946</td>
</tr>
<tr>
<td>C2-VD</td>
<td>$\mu_1$</td>
<td>0.993</td>
<td>1.044</td>
</tr>
<tr>
<td></td>
<td>$\mu_2$</td>
<td>2.988</td>
<td>2.200</td>
</tr>
<tr>
<td>C2-MVD</td>
<td>$\mu_1$</td>
<td>0.993</td>
<td>1.044</td>
</tr>
<tr>
<td></td>
<td>$\mu_2$</td>
<td>2.987</td>
<td>2.310</td>
</tr>
</tbody>
</table>

### E. Level noise and misclassification error

We consider different levels $(\Delta, \rho)$ of MD, VD and MVD noises. These noises are combined with C1 and C2. For these combinations, the behavior of the misclassification rate, computed by MPM method, using EV is given in Table VI. The performance of BEM and BBHMC through the following three cases $(\Delta, \rho) \in \{(5, 1), (0, 1), (0, 1.63)\}$ is presented in Fig. 1 to Fig. 6.

In each figure, we give the behavior of error rate (for each class) with respect to the iterations. In comparison with EV, we constat that BEM and BBHMC give good results for $(\Delta, \rho) = (5, 1)$ (see Fig. 1 and Fig.4); the error rates for BEM and BBHMC are nearer to those given by EV after $T_1 = 1000$ iterations for BBHMC and after 50 iterations for BEM. So BEM outperforms BBHMC in this easiest case (i.e. lower level
TABLE V
ESTIMATIONS OF $\Sigma_1$ AND $\Sigma_2$ BY EV, BEM AND BBHMC FOR DIFFERENT DATA SCENARIOS. THE STANDARD DEVIATION IS GIVEN BETWEEN BRACKETS FOR BBHMC ESTIMATIONS

<table>
<thead>
<tr>
<th>Data</th>
<th>$\Sigma_1$</th>
<th>$\Sigma_2$</th>
<th>BBHMC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_{11}$</td>
<td>$\sigma_{22}$</td>
<td>$\sigma_{12}$</td>
</tr>
<tr>
<td>C1+MD</td>
<td>1.990</td>
<td>3.362</td>
<td>0.279</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.019)</td>
<td>(0.316)</td>
</tr>
<tr>
<td>C1+VD</td>
<td>0.235</td>
<td>4.194</td>
<td>1.111</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.023)</td>
<td>(0.112)</td>
</tr>
<tr>
<td>C1+MVD</td>
<td>4.980</td>
<td>15.642</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>(5.006)</td>
<td>(15.698)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>C2+MD</td>
<td>4.309</td>
<td>13.823</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>(4.291)</td>
<td>(13.823)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>C2+VD</td>
<td>2.308</td>
<td>6.784</td>
<td>0.123</td>
</tr>
<tr>
<td></td>
<td>(1.992)</td>
<td>(6.756)</td>
<td>(0.123)</td>
</tr>
<tr>
<td>C2+MVD</td>
<td>4.980</td>
<td>15.642</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>(5.006)</td>
<td>(15.698)</td>
<td>(0.039)</td>
</tr>
</tbody>
</table>

Noise). In all other cases, BBHMC works better than BEM: the error rates, after stabilization, are nearer to those given by EV for BBHMC than for BEM (see Fig. 2, Fig. 3, Fig. 4 and Fig. 6).

TABLE VI
THEORETICAL ERROR RATES FOR EACH CLASS ($ER_1$ AND $ER_2$) FOR C1 AND C2. $ER_1$ IS THE MISCLASSIFICATION ERROR OF CLASS I.

<table>
<thead>
<tr>
<th>Markov chain</th>
<th>$\Delta = 5$</th>
<th>$\Delta = 0$</th>
<th>$\Delta = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 5$</td>
<td>$\rho = 1$</td>
<td>$\rho = 1$</td>
<td>$\rho = 1$</td>
</tr>
<tr>
<td>C1 $ER_1$</td>
<td>0.5</td>
<td>1.3</td>
<td>2.4</td>
</tr>
<tr>
<td>C1 $ER_2$</td>
<td>1.8</td>
<td>3.3</td>
<td>11.5</td>
</tr>
<tr>
<td>C2 $ER_1$</td>
<td>2.7</td>
<td>2.2</td>
<td>4.9</td>
</tr>
<tr>
<td>C2 $ER_2$</td>
<td>9.7</td>
<td>15.8</td>
<td>18.5</td>
</tr>
</tbody>
</table>

VI. UNSUPERVISED IMAGE SEGMENTATION
In the following, we compare BHMC using Hilbert-Peano scan and BHMC using a noisy supplementary information (for example: proportion of pixels in the same class as the current pixel in a neighborhood of the last) on image. We consider that a neighborhood of a pixel is composed of 3, 5 or 8 pixels around it according to the following three positions: corner, border or inside. The comparison is made through a sample of four synthetic images given in Fig. 7 (Alphabet, Gibbs, etc...)
The problem of this method is that the past and the future of a pixel doesn’t correspond always to its spatial context. Thus, when one considers the method “line by line”, two neighboring pixels and belonging to the same column are near spatially and distant in sense of Markov chain. To avoid the problem of the temporal and spatial contexts, a lot of authors use the Hilbert-Peano scan [6], [16], [19]. In this method, the noisy image is firstly transformed into one dimension chain using Hilbert-Peano scan and the segmented image is reconstructed using an inverse Hilbert-Peano scan. The proposed model BHMC using noisy supplementary information overcomes the problem of adequacy between temporal and spatial context without using the Hilbert-Peano scan.

### A. BHMC without and with Hilbert-Peano scan

The simplest idea for modeling an image by Markov chain is to consider this image “line by line” or “column by column”.

### B. BEM and BBHMC performances

Before analyzing the results, let us note that for each statistical image segmentation method, there exists a theoretical error, obtained with the true parameter values [3]. In our case this error is estimated from empirical values EV (i.e. obtained from real and noisy images). For a simplification reason, we just interest to the error rates (wrongly classified pixels) and we do not give the parameters estimation. Thus we present,

![Image](image-url)
in Table IX, the error rates for each class (ER_1 and ER_2) obtained by EV, BEM and BBHMC on corrupted images. 
We constat that BEM and BBHMC, compared to EV, prove successful in image Letter B. This can be explained by the fact that the two pixel classes are well separated. Also we constat that BEM is better than EV and BBHMC for class 2 in all cases, but the mean error rates given by BBHMC are significantly reduced than those obtained by BEM and they are nearer to those given by EV. The mean reduction factors between BBHMC and BEM are 2.65, 1.39, 1.79 and 1.53 for Alphabet, Gibbs, Letter B and Ring respectively. All the estimations, for each data scenario, are obtained by BEM and BBHMC for a mean CPU time equal to 36.5.

VII. CONCLUSION

We have proposed a bi-dimensional Markov chain model, BHMC, mathematically designed by \((X, Y)\) where \(X\) is the unobserved Markov chain and \(Y = (Y^1, Y^2)\) is the observed process. In previous works \(Y^1\) and \(Y^2\) represent two images when, in our case \(Y^1\) represent noisy image and \(Y^2\) represent noisy supplementary information on the studied image (we have taken \(Y^2\) as proportion of pixels in the same class as the current pixel in a neighborhood of the last). For the proposed model, we have considered a bayesian algorithm BBHMC to estimate the model parameters. We have compared BBHMC to the classical Expectation-Maximization algorithm in bi-dimensional case, BEM. The results of these two algorithms are evaluated by empirical values, EV, computed from complete data. Our propositions are favorably tested on simulation and real data. We have considered different data scenarios and we have obtained satisfactory results for our approach: the obtained parameters estimation are nearer to those given by EV and the mean factor reduction of error rates with respect to BEM is between 1.33 and 2.65 for image segmentation. Also the proposed model BBHMC outperforms (see Tables VII and VIII) the same model considered with Hilbert-Peano scan.

Actually, we interest to the generalization of this work to multi-dimensional case by integrating supplementary information to each image \((i.e. Y = (Y^1, Y^1, Y^2, Y^2, \ldots, Y^M, Y^M))\), where \(Y^i\) is an image and \(Y^1i\) is a supplementary information on \(Y^2\), for \(i = 1, \ldots, M\) with \(M > 2\).


### REFERENCES


