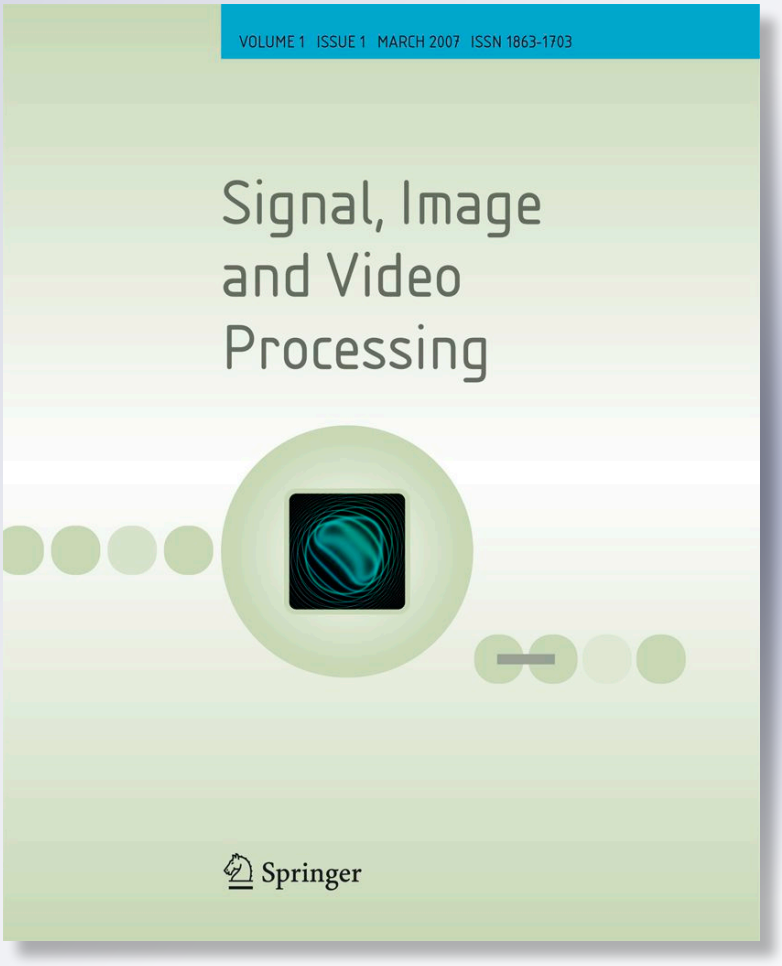
Subsampling-based HMC parameter estimation with application to large datasets classification

Stéphane Derrode, Lamia Benyoussef & Wojciech Pieczynski

Signal, Image and Video Processing

ISSN 1863-1703

SIViP
DOI 10.1007/s11760-012-0324-2



Springer

SIViP
DOI 10.1007/s11760-012-0324-2

ORIGINAL PAPER

Subsampling-based HMC parameter estimation with application to large datasets classification

Stéphane Derrode · Lamia Benyoussef · Wojciech Pieczynski

Received: 4 January 2012 / Revised: 12 April 2012 / Accepted: 13 April 2012

Abstract This paper presents a contextual algorithm for the approximation of Baum's forward and backward probabilities, which are extensively used in the framework of Hidden Markov chain models for parameter estimation. The method differs from the original algorithm by taking into account only a neighborhood of limited length and not all the data in the chain for computations. It then becomes possible to propose a bootstrap subsampling strategy for the computation of forward and backward probabilities, which greatly reduces computation time and memory saving required for EM-based parameter estimation. Comparative experiments regarding the neighborhood size and the bootstrap sample size are conducted by mean of unsupervised classification error rates. Practical interest of such an algorithm is then illustrated through the segmentation of large-size images; classification results confirm the validity and the accuracy of the proposed algorithm while greatly reducing computation and memory requirements.

Keywords Hidden Markov Chain · Estimation–maximization · Forward and backward probabilities · Bootstrap resampling · Image segmentation

S. Derrode (✉) · L. Benyoussef
Institut Fresnel (CNRS UMR 7249), Universités de Marseille and École Centrale Marseille, 38, rue Frédéric Joliot-Curie, 13451 Marseille Cedex 20, France
e-mail: stephane.derrode@centrale-marseille.fr

W. Pieczynski
CITI Department (CNRS UMR 5157), TELECOM SudParis, 9, Rue Charles Fourier, 91011 Evry Cedex, France
e-mail: Wojciech.Pieczynski@it-sudparis.eu

1 Introduction

The discrete Hidden Markov chain (HMC) model has shown to be very performing in a number of different contexts, covering economical prediction, health sciences and, especially, signal and image processing, with applications in automatic speech recognition [25], gesture analysis [28], and image segmentation [3], for examples. From its first introduction in the seventies, a great amount of effort has been devoted to the development and improvement of the HMC structure at several levels. Indeed, the HMC model has been extended with different topologies including pairwise [10], triplet [24], and higher-order [4] HMCs, and with other architectures like HMM2 [27], factorial [16], and coupled [29] HMCs, among others. All those methods improve the capabilities of the classic HMC to model more complex situations (correlated noise, coupled data, …), but to the cost of an increase of algorithms complexity and memory requirements when model parameters have to be estimated.

In an unsupervised context, HMC model parameters need to be estimated from available data before Bayesian classification, either using EM (Expectation–Maximization) [9], SEM (Stochastic EM) [7], or ICE (Iterative Conditional Estimation) [22] procedures. Such algorithms become very time and memory demanding in case of huge datasets like network traffic data or large satellite images for which size can exceed 10^8 pixels. One solution proposed in [6] is to divide data into overlapping windows of fixed size (say 100 pixels length for example) and to classify the central data of each window using parameters estimated on the window only. However, this sliding window method faces the following problem. Suppose that the entire dataset must be classified in two classes. All windows do not necessarily contain data from the two classes but can contain data from one class only (think about an homogeneous region in an image for example). This leads

Published online: 10 May 2012

to estimation problems. To remedy, authors proposed to estimate the number of classes in each window using Bayesian or Akaike Information Criteria (BIC and AIC).

One main reason for time and memory consumption of the HMC model comes from the computation and saving of Baum's *Forward* and *Backward* (F&B) probabilities [2], which are extensively used for unsupervised parameter estimation (whatever the estimation procedure). Especially, the memory requirements needed to save those probabilities can become prohibitive for large-size datasets or when the number of classes is high. Since data collections tend to grow, it is interesting to propose a solution that preserves the nice properties of the HMC model which as proved itself to be very effective in a number of situations, especially in case of strongly noisy signals [15]. In this work, we propose a contextual algorithm for the computation of F&B probabilities. The algorithm differs from the original one since it takes into account a neighborhood of limited extent and not all the chain for probability computations. As explained latter in details, the approach is justified by the short memory property of Markov processes, that is, the correlation between data "decreases quickly to zero" as the distance between them increases. It then becomes possible to select a bootstrap representative subsample of data, for which F&B probabilities are computed. By adapting the original EM procedure to this resampling context, the estimation of HMC parameters is performed on the selected dataset only, which greatly reduces computing time and memory requirements.

The paper is organized as follows. Section 2 recalls basic facts about HMC-based Bayesian classification and unsupervised parameter estimation using EM procedure. Section 3 presents the main novelties of the paper, that is, a new contextual algorithm for the approximation of F&B probabilities and its integration in a bootstrap subsampling scheme for fast unsupervised parameter estimation. Performance of algorithms is systematically evaluated on numerous experiments using synthetic noisy data. Finally, algorithms are illustrated in the context of large-size images segmentation in Sect. 4. Conclusion and further work are drawn in Sect. 5.

2 HMC model and EM-based estimation

This section is devoted to set notations and to recall some basics facts about EM-based parameter estimation in the HMC context.

2.1 Hidden Markov chain model

Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ be two random processes corresponding to the (hidden) state sequence and the observed sequence. N denotes the number of observations to be classified. Each random variable X_n takes its values in a finite set of classes $\Omega = \{1, \ldots, K\}$, and each random variable Y_n takes its values in the set of real numbers $\mathbb{R}$. The problem of estimating $\boldsymbol{X}$ from $\boldsymbol{Y}$, which occurs in numerous applications, can be solved with Bayesian methods once one has chosen some accurate distribution $p(\boldsymbol{x}, \boldsymbol{y})$ for $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$.

For time series, the hidden Markov chain model is the simplest and most well-known model for $\boldsymbol{Z}$ [14]. In this model, $\boldsymbol{X}$ is assumed to be a homogeneous Markov chain, that is, the entries of the $\boldsymbol{C}$ matrix $c(i, j) = p(X_n = i, X_{n+1} = j)$, for all $i, j \in \Omega$, are independent of n. The distribution of $\boldsymbol{X}$ is consequently determined by an initial distribution vector, denoted by $\boldsymbol{\pi}$, which entries are defined by

$$\forall i \in \Omega, \quad \pi(i) = p(X_n = i) = \sum_{l=1}^{K} c(i, l) \tag{1}$$

and a transition matrix $\boldsymbol{T}$ which entries are given by:

$$\begin{aligned} \forall i, j \in \Omega, \quad t(i, j) &= p(X_{n+1} = j \,|X_n = i) \\ &= \frac{c(i, j)}{\pi(i)} \end{aligned} \tag{2}$$

Following usual assumptions

- random variables $Y_1, \ldots, Y_N$ are conditionally independent with respect to $\boldsymbol{X}$, and
- the distribution of each Y_n conditionally on $\boldsymbol{X}$ is equal to its distribution conditional on X_n,

we write $p(\boldsymbol{y}\,|\boldsymbol{x}) = \prod_{n=1}^{N} p(y_n\,|x_n)$. The K data-driven densities $p(y_n\,|x_n) = f_{x_n}(y_n)$ model the "noise" or the variability of each class.

Hence, the distribution of $\boldsymbol{Z}$ writes

$$p(\boldsymbol{x}, \boldsymbol{y}) = \pi(x_1)\, f_{x_1}(y_1) \prod_{n=2}^{N} t(x_{n-1}, x_n)\, f_{x_n}(y_n)$$

It is well-known that process $\boldsymbol{X}\,|\boldsymbol{Y}$ is a Markov chain and Bayesian restoration[1] is made possible according to the MAP (Maximum A Posteriori) criterion

$$\widehat{\boldsymbol{x}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{x} \in \Omega^N} p(\boldsymbol{x}\,|\boldsymbol{y}) \tag{3}$$

using Viterbi's algorithm [25], or the MPM (Maximum Posterior Mode) criterion

$$\widehat{\boldsymbol{x}}_{\mathrm{MPM}}(\boldsymbol{y}) = (\widehat{x}_1, \ldots, \widehat{x}_N) \tag{4}$$

with $\widehat{x}_n = \arg\max_{x_n \in \Omega} p(x_n\,|\boldsymbol{y})$

Bayesian classification requires the entries of the $\boldsymbol{C}$ matrix and the set $\boldsymbol{\Theta}$ of parameters for the K data-driven densities,

[1] that is, minimum mean error rate-based restoration

for example, means and standard deviations when assuming Gaussian distributions. Most of the time, these parameters are not known and must be estimated from observations only. This can be achieved using EM [5,9,20], SEM [7], or ICE [8,22,23] iterative procedures. The first two methods are based on likelihood maximization, whereas the third one is based on conditional expectation of complete data estimators in a mean-square error sense. SEM and ICE require the simulation of $\boldsymbol{X}\,|\boldsymbol{Y}$, which cannot be adapted to the context described here (see remark 2) so that we only consider EM-based estimation.

2.2 EM-based parameter estimation

One of the nice properties of HMC we are interested in this paper is that all the posterior distributions $p\,(X_n = k\,|\boldsymbol{y})$ and $p\,(X_n = l, X_{n+1} = k\,|\boldsymbol{y})$ are calculable, even for large N. The method is based on the so-called Baum's forward and backward recursion algorithm [2,11,20] and recalled in the sequel for latter use. According to the synoptic sketched in Fig. 1, all probabilities presented below are computed at each iteration EM $\ell \in [1, \mathcal{L}]$ and notations should depend on ℓ. However, we omit reference to ℓ for clarity and simplicity.

The forward probabilities are defined by

$$\alpha_n(k) = p\,(X_n = k\,|y_{1:n})$$

with

$$\alpha_1(k) = \frac{\pi(k)\ f_k\,(y_1)}{\sum_{l=1}^{K} \pi(l)\ f_l\,(y_1)}$$

$$\alpha_n(k) = \frac{1}{S_n}\ p\,(X_n = k,\, y_n\,|y_{1:n-1}) \tag{5}$$

and can be computed using the following recursion

$$\alpha_{n+1}(k) = \frac{1}{S_{n+1}}\ f_k\,(y_{n+1})\ \sum_{l=1}^{K} t(l,k)\ \alpha_n(l), \tag{6}$$

with $S_1 = p\,(y_1)$ and for $1 \leq n \leq N$, $S_n = p\,(y_n\,|y_{1:n-1})$. The backward probabilities are defined by $\beta_N(k) = 1$ and, $\forall n < N$

$$\begin{aligned}\beta_n(k) &= \frac{p\,(y_{n+1:N}\,|X_n = k)}{p\,(y_{n+1:N}\,|y_{1:n})} \\ &= \frac{1}{S_{n+1}}\ \frac{p\,(y_{n+1:N}\,|X_n = k)}{p\,(y_{n+2:N}\,|y_{1:n+1})}\end{aligned} \tag{7}$$

They can also be computed recursively using

$$\beta_n(k) = \frac{1}{S_{n+1}}\ \sum_{l=1}^{K} t(k,l)\ f_l\,(y_{n+1})\ \beta_{n+1}(l) \tag{8}$$

for all $k \in \Omega$

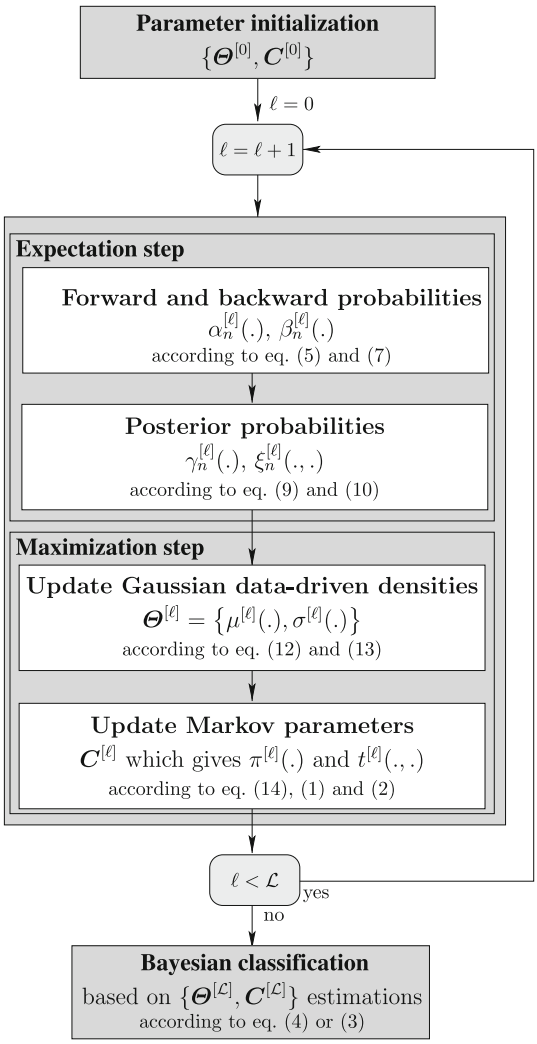


Fig. 1 Flowchart for unsupervised HMC classification: initialization, EM-based estimation, and classification. ℓ denotes the iteration number and $\mathcal{L}$ the number of iterations to reach convergence

Then, it can easily be shown that *a posteriori* marginal and joint distributions can be expressed as follows:

$$\gamma_n(k) = p\,(X_n = k\,|\boldsymbol{y}) = \alpha_n(k)\ \beta_n(k) \tag{9}$$

$$\begin{aligned}\xi_n(l,k) &= p\,(X_n = l,\, X_{n+1} = k\,|\boldsymbol{y}) \\ &= \frac{\alpha_n(l)\ \beta_{n+1}(k)\ t(l,k)\ f_k\,(y_{n+1})}{S_{n+1}}\end{aligned} \tag{10}$$

Remark 1 Due to the sum-to-one property of forward probabilities in (5), the normalization coefficients write

$$S_{n+1} = \sum_{k=1}^{K} f_k\,(y_{n+1})\ \sum_{l=1}^{K} t(l,k)\ \alpha_n(l). \tag{11}$$

At iteration ℓ, re-estimation formulae for mean and variance of Gaussian data-driven densities are given by

$$\mu^{[\ell]}(k) = \frac{\sum_{n=1}^{N} \gamma_n^{[\ell]}(k)\ y_n}{\sum_{n=1}^{N} \gamma_n^{[\ell]}(k)}, \tag{12}$$

$$\sigma^{2\,[\ell]}(k) = \frac{\sum_{n=1}^{N} \gamma_n^{[\ell]}(k)\ \left(y_n - \mu^{[\ell]}(k)\right)^2}{\sum_{n=1}^{N} \gamma_n^{[\ell]}(k)}, \tag{13}$$

while re-estimation formula for $\boldsymbol{C}$ matrix entries is given by

$$c^{[\ell]}(l,k) = \frac{1}{N-1} \sum_{n=1}^{N-1} \xi_n^{[\ell]}(l,k). \tag{14}$$

A priori and transition probabilities are obtained using Eqs. (1) and (2), respectively.

To start the iterative procedure, good initial value is required for each parameter. For all experiments presented below, we used the k means segmentation algorithm with K classes and classical estimators from complete data.

3 Resampling-based EM parameter estimation

The section intends to present an original algorithm for quick but accurate approximation of HMC model parameters based on EM principle. First, a contextual approximation of forward and backward probabilities is presented and evaluated through a set of experiments on synthetic noisy data. Then, the algorithm is integrated in a subsampling scheme to make HMC parameter estimation fast and memory saving, while keeping accuracy as confirmed latter by extensive experiments.

3.1 Contextual approximation of forward and backward probabilities

The gradual computation of forward and backward probabilities illustrates the global behavior of Markov chain modeling, that is, each data of a time series are linked to all previous data by the one just before. Given the n^{th} data, one can ask for the influence of a data with index $m < n$ on the computation of $\alpha_n(.)$, knowing that a Markov chain is a "short memory" process. The same question arises for $\beta_n(.)$.

To evaluate such an influence, the following experiment has been conducted: for each observation, the forward and backward recursion rules are applied only considering a limited number of neighboring data around it:

- Extract a subchain around the data with index n delimited by range $[n-\lambda; n+\lambda]$, see Fig. 2. The window size is then $2\lambda+1$.
- Apply the forward and backward recursion algorithms [see Eqs. (5) and (7)] considering the subprocesses $\widetilde{\boldsymbol{X}}_{1:2\lambda+1} = \boldsymbol{X}_{n-\lambda:n+\lambda}$ and $\widetilde{\boldsymbol{Y}}_{1:2\lambda+1} = \boldsymbol{Y}_{n-\lambda:n+\lambda}$.

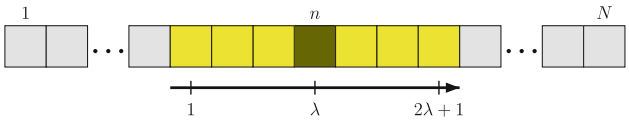


Fig. 2 Neighborhood used to estimate α_n and β_n

- Store probabilities $\alpha_n(k) = \tilde{\alpha}_{\lambda+1}(k)$ and $\beta_n(k) = \tilde{\beta}_{\lambda+1}(k)$ for all $k \in \Omega$ and discard all other probabilities.

These steps are reproduced for all data of index $n \in [1, N]$, with special care on the first and last λ observations. Hence, this way we get a contextual approximation of forward and backward probabilities for all the data. The next steps to parameter estimation remain the same (see Fig. 1).

The influence of the neighborhood on the algorithm accuracy is now evaluated with respect to λ, by means of classification error rates computed on noisy synthetic data. To simulate noisy data, we first simulate $N = 65536$ samples (which corresponds to an image with size 256×256) following a $K = 2$ classes Markov chain model ($\boldsymbol{x}$) and then add some Gaussian noise to each class to get a set of N noisy data ($\boldsymbol{y}$). More precisely, parameters used for data generation are:

- Markov chain parameters: $\boldsymbol{\pi} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, $\boldsymbol{T} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$.
- Gaussian data-driven densities: means $\mu_1 = 0$ and $\mu_2 = 1$; standard deviations $\sigma_1 = \sigma_2 = 1$.

For each experiment, the number of iterations for EM was set to $\mathcal{L} = 60$, assuming convergence. The simulated Markov chain $\boldsymbol{x}$ is used as a ground truth for comparison with unsupervised restorations obtained from $\boldsymbol{y}$ only. The MPM-based classification results are means of 50 independent experiments.

The supervised restoration of $\boldsymbol{y}$[2] gives a mean error rate of 17.85% (std: 0.28). The unsupervised restoration of $\boldsymbol{y}$ with parameters estimated using the classical algorithm described in Sect. 2 gives a mean error rate of 17.86% (std: 0.27). These results confirm the very nice behavior of EM in the context considered here.

Unsupervised restoration results of $\boldsymbol{y}$ are reported in Fig. 3, with parameters estimated using the algorithm described above for increasing values of λ. As expected, the error rate decreases when the size of the window increases. Indeed, as λ grown, we get a better approximation of forward and backward probabilities, and so a better estimation of model parameters used for Bayesian classification. In this experiment, the supervised error rate is reached for $\lambda \leq 3$.

[2] that is, restoration with true parameters

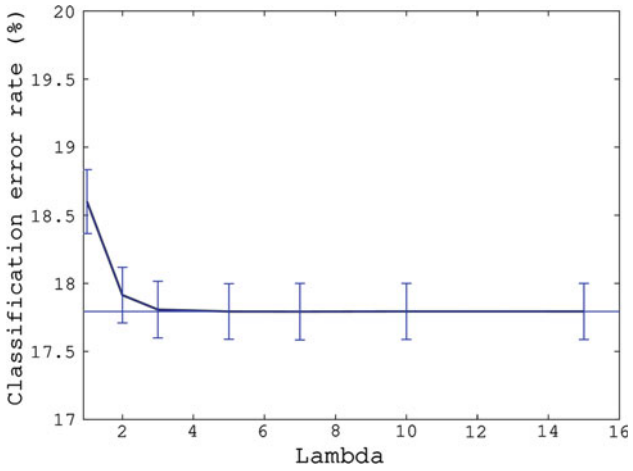


Fig. 3 Classification error rates versus the neighbor size (λ) for MPM classification, when model parameters are estimated using the contextual approximation proposed

This experiment confirms that the influence of far apart data decreases rapidly on forward and backward probability computations, as a consequence of the so-called "short memory" property of Markov processes. Next subsection tries to follow up this behavior to design a fast and accurate algorithm for EM-based parameter estimation. It should be, however, noted that the λ value is application dependent and should be carefully chosen. For example, in speech modeling applications, the structure of HMM is generally left-to-right and the number of states very large (a few dozens). In that case, the context must be large enough to avoid confusion between phonemes and words in recognition.

3.2 Subsampling-based parameter estimation

Even if forward and backward recursion algorithms make posterior probabilities $\gamma_n(.)$ and $\xi_n(.,.)$ computable for large N, computer memory requirements for storing all probabilities limit its usage in practice. In satellite image segmentation for example, we more and more face the problem of segmenting image with more than 10^8 pixels. Furthermore, since computation time is linear with respect to N, estimation time can be prohibitive for such datasets. We now propose to combine the algorithm described in previous section to estimate parameters with a representative subset of the original dataset.

According to previous experiments, it is possible to get a good approximation of forward and backward probabilities at index n without requiring to compute $\alpha_{n-1}(.)$ and $\beta_{n+1}(.)$ (if λ is large enough). The main idea the proposed algorithm relies on is to estimate HMC parameters based only on a subsample $\boldsymbol{w} = \{w_1, \ldots, w_M\}$ of the original dataset $\boldsymbol{y} = \{y_1, \ldots, y_N\}$, with $M \ll N$.

The choice of an optimal representative subset $\mathcal{A}$ is crucial, and we make use of the bootstrap resampling technique as proposed by Efron [12,13]. A bootstrap sample is obtained by independent drawings with replacement from the empirical distribution (given by data histogram H). The question of finding an optimal number $M = \text{card}\,\{\mathcal{A}\}$ of representative data has been addressed in [1] (see also [21]): if G denotes the number of bins in H then M is the highest value such that $T(M) < \epsilon$ with

$$T(M) = \sum_{g=1}^{G} \frac{H(g)\, e^{-MH(g)}}{1 - e^{-MH(g)}}$$

and ϵ a fixed small value set to 0.03 for image segmentation experiments.

Parameters re-estimation formulae in Eq. (12) to (14) rewrite

$$\mu^{[\ell]}(k) = \frac{\sum_{a\in\mathcal{A}} \gamma_a^{[\ell]}(k)\; y_a}{\sum_{a\in\mathcal{A}} \gamma_a^{[\ell]}(k)}, \tag{15}$$

$$\sigma^{2\,[\ell]}(k) = \frac{\sum_{a\in\mathcal{A}} \gamma_a^{[\ell]}(k)\; \left(y_a - \mu^{[\ell]}(k)\right)^2}{\sum_{a\in\mathcal{A}} \gamma_a^{[\ell]}(k)}, \tag{16}$$

$$c^{[\ell]}(l,k) = \frac{1}{M_1} \sum_{a\in\mathcal{A}} \xi_a^{[\ell]}(l,k), \tag{17}$$

with

$$\gamma_a^{[\ell]}(k) = \alpha_a^{[\ell]}(k)\; \beta_a^{[\ell]}(k), \tag{18}$$

$$\xi_a^{[\ell]}(l,k) = \frac{\alpha_a^{[\ell]}(l)\; \beta_{a+1}^{[\ell]}(k)\; t^{[\ell]}(l,k)\; f_k^{[\ell]}(y_{a+1})}{S_{a+1}^{[\ell]}}. \tag{19}$$

for all $a \in \mathcal{A}$. Probabilities$\alpha_a^{[\ell]}(.)$ and $\beta_a^{[\ell]}(.)$ are computed at each EM iteration ℓ using the contextual approximation presented in previous subsection. Regarding Eq. (19), $S_{a+1}^{[\ell]}$ can be computed using (11) that only requires $\alpha_a^{[\ell]}(.)$. The computation of $\beta_{a+1}^{[\ell]}(.)$ can be easily obtained by solving the square system in Eq. (8), which only requires $\beta_a^{[\ell]}(.)$.

To illustrate the algorithm, we continued experiments reported before (all parameter values remain). From each of the 50 simulated samples, we extracted a representative subset of size M that is used for parameter estimation according to the algorithm detailed above. More precisely, the subsample is redrawn at each EM iteration to avoid degenerate behaviors. To measure performances, we have studied the influence of λ and M on misclassification rates. Results are reported in plots of Fig. 4. As expected, whatever the subsample size, error rates decrease as λ increases. It can be observed that, for a value of λ between 5 and 10, the error rates reach a constant value. For low sample sizes, classification results show a higher variability as suggested by the increase of the standard deviation of error rates.

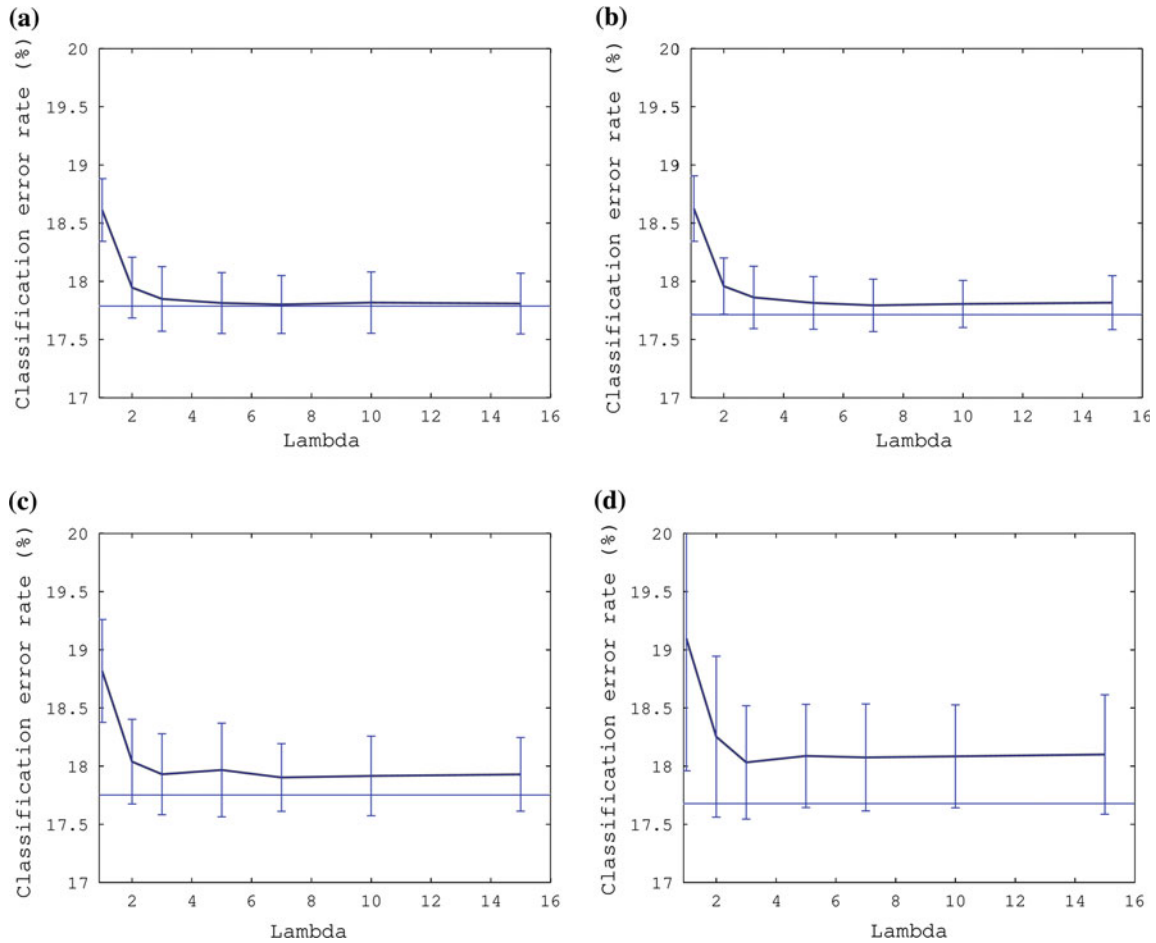


Fig. 4 Error rates versus the neighbor size (λ) when model parameters are estimated using the contextual approximation with varying subsample sizes. **a** $M = 14288$ (21.80 %), **b** $M = 3368$ (5.13 %), **c** $M = 2003$ (3.06 %), **d** $M = 911$ (1.39 %),

Mean computing time for all the experiments is reported in plots of Fig. 5. The horizontal line represents the processing time required by the classical EM estimation procedure (41 s) for $\mathcal{L} = 100$ iterations. The colored lines represent the computing time for the same experiment but considering the subsampling approximation strategy described in this work, with varying number of samples (from $M = 911$ to 14, 288) and varying window sizes (from $\lambda = 1$ to 40). As expected, the computing time is linear with respect to λ. It is also linear with respect to the number of samples selected for parameter estimation. One can note that the gain in computational efficiency is not given by the ratio of M by N. The reason is twofold:

- The estimation of α_a and β_a for the sub-sample is more time-consuming than the estimation of α_n and β_n for the original sample. Indeed, the estimation of α_a and β_a requires the computation of $2\lambda + 1$ temporary forward and backward probabilities.
- The time required to sub-sample the data at each iteration of EM is not negligible.

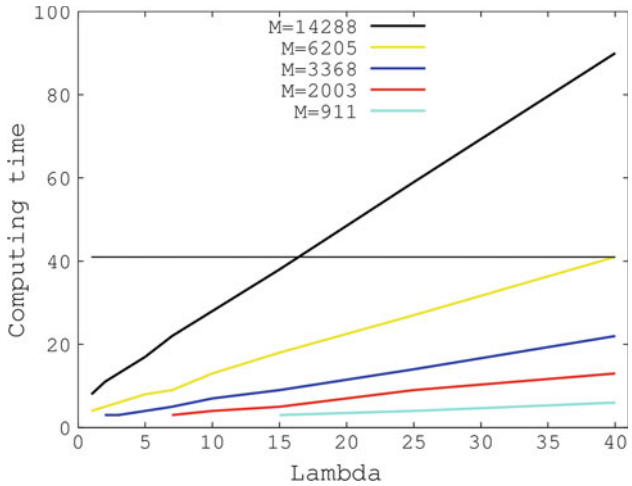


Fig. 5 Processing time for four subsample sizes M with respect to the neighbor size λ. The *horizontal line* denotes the computing time required by the classical EM procedure

Remark 2 Unlike EM, SEM and ICE estimation procedures require simulations of $\boldsymbol{X}$ conditionally to $\boldsymbol{Y}$, which is

a non-homogeneous Markov chain with initial law given by $p(X_1 = k \,|\, \mathbf{y}) = \gamma_1(k)$ and transition matrix given by

$$\tilde{a}_n(l,k) = p(X_{n+1} = k \,|\, X_n = l, \mathbf{y}) = \frac{t(l,k)\ f_k(y_{n+1})\ \beta_{n+1}(k)}{S_{n+1}\ \beta_n(l)}.$$

Hence, the simulation of x_n requires x_{n-1} which is not known in the contextual approximation proposed here. A solution is to simulate $\boldsymbol{x}$ *a posteriori* according to its marginal distribution given by $\gamma_n(.)$, as experimented in [3].

4 Application to large-size image segmentation

Experiments described in the previous section have been conducted on noisy simulated Markov chains. We can now wonder whether the same favorable behavior can be observe when considering real data, that is, data that are not supposed to follow a Markov chain model, such as image data. Whatever they are photographic, medical, or satellite based, images reach millions of pixels due to technological improvements in sensor resolution, which make them good candidates for the proposed algorithm.

Hence, we propose to evaluate the contextual algorithm to segment images according to the HMC model, following works in [3,15,17]. Here are the main steps to process an image with the HMC model

- The bi-dimensional lattice of pixels is first converted into a 1D sequence of observations ($\mathbf{y}$) through the Hilbert-Peano scan [26].
- Then, parameter estimation and Bayesian restoration techniques can be applied to obtain a restored sequence of class data ($\mathbf{x}$),
- Finally, $\mathbf{x}$ is converted back to a class image by inverting the scan [3].

We only select a small representative sample of pixels in the generated Peano sequence to estimate parameters using the contextual algorithm described in Sect. 3, see Fig. 6.

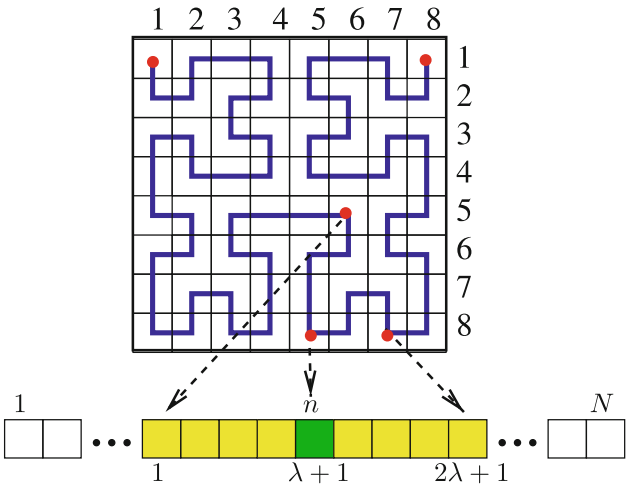


Fig. 6 Illustration of the Peano scan on a 8×8 image and of the approximation algorithm on the generated sequence (*red dots* represent bootstrap subsample) (color figure online)

4.1 Experiment on a photographic image

The algorithm has first been evaluated on the 4000×2000 photographic image in Fig. 7. The image is noisy due to under-illumination at the time of snapshot. The image was segmented using the classical EM-based parameter estimation and the contextual approximation described in this paper. The image size is the maximum allowed for the computer we used. We set $K = 4$ and $\mathcal{L} = 100$. For the contextual algorithm, we set $\lambda = 3$ and $M = 1308$ pixels, which represents $1.67^0/_{000}$ of the total number of pixels. Estimated HMC model parameters for the two algorithms have been reported in Table 1. As can be seen, parameter values are very close to

Fig. 7 Original photographic image with 8 million pixels, showing part of the town of Marseilles, France. The rectangle delimits the area for which segmentation results is compared in Fig. 8

Table 1 Parameters estimated to segment image in Fig. 7 with (a) the classical EM estimation procedure, and (b) the contextual estimation procedure

	$\boldsymbol{\pi}$	$\boldsymbol{T}$	$(\mu_k, \sigma_k^2)_{k\in[1,4]}$
(a)	$\begin{pmatrix}0.32\\0.22\\0.25\\0.21\end{pmatrix}$	$\begin{pmatrix}0.99 & 0.01 & 0.00 & 0.00\\0.01 & 0.96 & 0.02 & 0.00\\0.00 & 0.02 & 0.97 & 0.01\\0.00 & 0.00 & 0.02 & 0.98\end{pmatrix}$	(48.6, 197.1) (95.4, 215.5) (134.0, 135.4) (170.6, 187.7)
(b)	$\begin{pmatrix}0.31\\0.22\\0.26\\0.21\end{pmatrix}$	$\begin{pmatrix}0.99 & 0.01 & 0.00 & 0.00\\0.01 & 0.97 & 0.02 & 0.00\\0.00 & 0.02 & 0.97 & 0.01\\0.00 & 0.00 & 0.01 & 0.99\end{pmatrix}$	(48.0, 195.9) (94.7, 203.8) (133.8, 127.7) (169.5, 163.0)

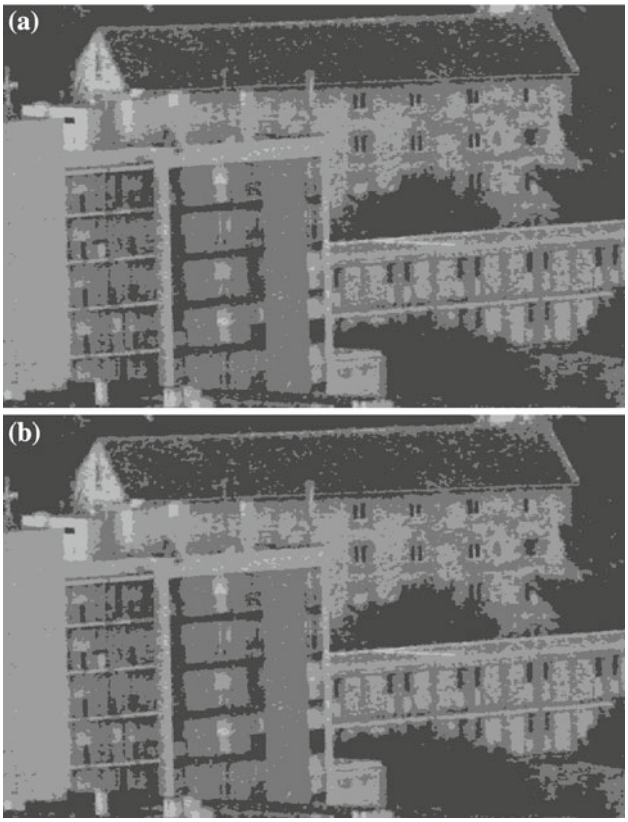


Fig. 8 Crop (1,000 × 650) of 4-classes segmentations for image in Fig. 7 using **a** the classical estimation procedure and **b** the contextual estimation procedure ($K = 4$, $\mathcal{L} = 100$, $\lambda = 3$ and $M = 1{,}308$). Class values (between 0 and 3) have been replaced by the mean gray-value of each class, see Table 1

each other, leading to nearly identical segmentation results, see Fig. 8. The time spent to estimate parameters with the classical algorithm is 3 h and 3 min and only 17 min and 45 s for the contextual approximation.

4.2 Experiment on a spot V image

The algorithm has also been evaluated on the 3000 × 3000 Spot V satellite image in Fig. 9, over the Arcachon basin, located on the French Atlantic coast. We set $K = 5, \mathcal{L} = 100$,

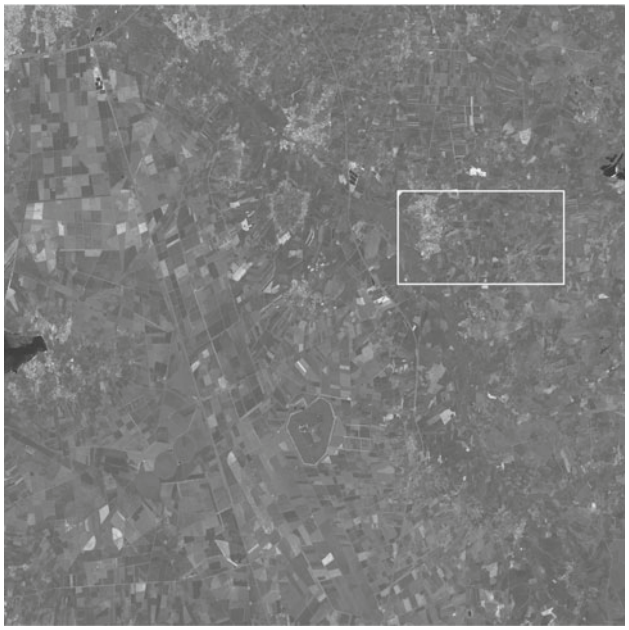


Fig. 9 Spot V satellite image, French Arcachon basin, ©CNES 2010—Distribution Spot Image. The rectangle (800 × 450) delimits the area for which segmentation results will be displayed for comparison

$\lambda = 3$, and $M = 3,248$ pixels, which represents $3.6^0/_{000}$ of the total number of pixels. The segmented image is reported in Fig. 10c and can be compared to the results obtained with two classical "blind-based" algorithms in Fig. 10a, b. Obviously one can observe the stronger regularization effect in segmented image obtained from the HMC model. What is important to note is that the contextual algorithm allows to segment large images that the classical algorithm cannot deal with.

If we account for the double precision (i.e., 8 bytes) arrays the classical and proposed algorithms require to store the numerous probabilities involved in EM (forward, backward, and marginal a posteriori), the peak memory consumption reaches the following quantities:

- classical algorithm: $N(3K + 1)$ * 8 bytes
- proposed algorithm: $M(3K + 1) + 2(2\lambda + 1)K$ * 8 bytes $\approx M(3K + 1)$ * 8 bytes

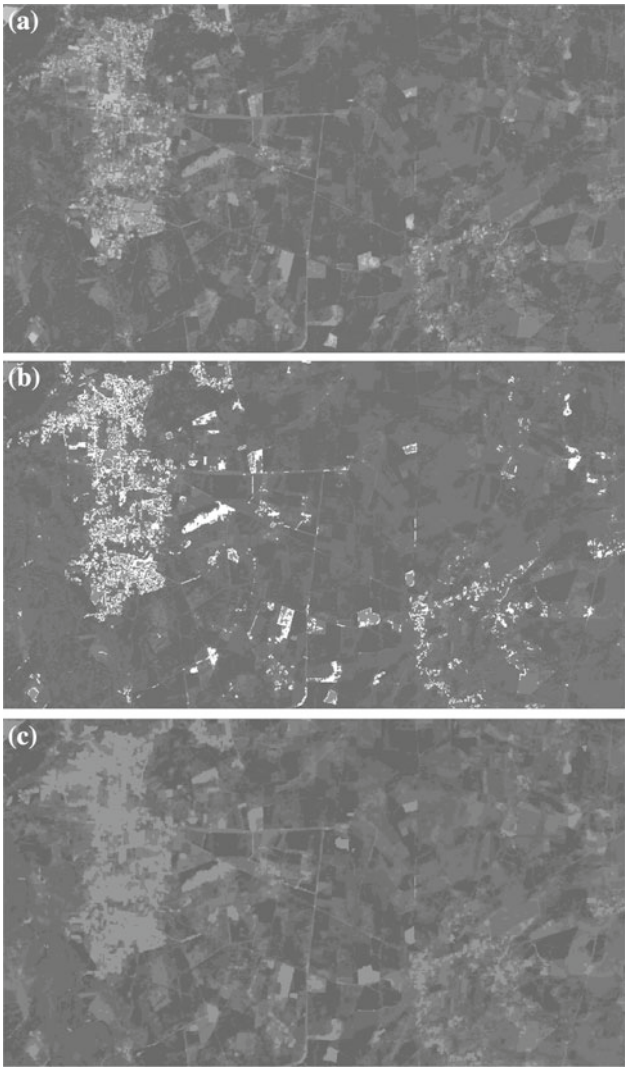


Fig. 10 Crop (800×450) of 5-classes segmentations for image in Fig. 9 using **a** the k means algorithm and **b** an EM-based "blind" Bayesian classification procedure ($K = 5$ and $\mathcal{L} = 100$), and **c** the contextual estimation procedure ($K = 5$, $\mathcal{L} = 100$, $\lambda = 3$ and $M = 3{,}248$). Class values (between 0 and 4) have been expanded to cover the range [0, 255] for visualization

so that the memory requirement gain is given by the ratio between N and M, which is equal to 2,770 for the Spot image of Arcachon basin.

5 Conclusion

This work describes an algorithm for the contextual estimation of forward and backward probabilities, and its application for EM-based unsupervised segmentation of large datasets using the HMC model. This approximation consists in considering only a neighborhood of limited extent in the computation of probabilities, and not all the chain as it is done in the original algorithm. The proposed method preserves the global property of the HMC model (still governed by only one set of parameters) while offering a contextual approximation. Then, a bootstrap subsampling strategy, which takes benefit of the previous algorithm, is proposed to get a quick but accurate estimation of model parameters based on EM. We validate our approach through a set of experiments on synthetic data and we show that the use of Bootstrap resampling can reach similar level of accuracy and robustness as the basic algorithm, yet it amounts to a considerable processing speed up. It is, however, important to note that the size of the context (λ) is application dependent and should be carefully chosen.

Experiments on large-size photographic and satellite images are used to illustrate the algorithm in real situations, where the original algorithm cannot be applied due to its computational and memory requirements. Of course, the algorithm is not restricted to deal with image data but can be of interest in any situation where data size to be explored is huge.

This contextual algorithm can be adapted to forward- and backward-like probabilities defined for the recent pairwise [10] and triplet [18,19] Markov chain models, which are strictly more general than the HMC model studied here. Also, in this work, we focused on scalar data and Gaussian mixtures for sake of clarity, but the algorithm can be extended to deal with generalized mixtures and multi-sensor data (e.g., in multi-spectral imagery).

Acknowledgments Authors would like to thank French *Centre National d'Études Spatiales* for providing Spot V images used in experiments.

References

1. Banga, C., Ghorbel, F.: Optimal bootstrap sampling for fast image segmentation: application to retina image. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'93), pp. 638–641. Minneapolis, MN (1993)
2. Baum, L., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. **41**, 164–171 (1970)
3. Benmiloud, B., Pieczynski, W.: Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images. Traitement du Signal **12**(5), 433–454 (1995). in French
4. Benyoussef, L., Carincotte, C., Derrode, S.: Extension of higher-order HMC modeling with application to image segmentation. Digit. Signal Process. **18**(5), 849–860 (2008)
5. Bilmes, J.: A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical report icsi-tr-97-021, University of Berkeley (1997)
6. Bouyahia, Z., Benyoussef, L., Derrode, S.: Change detection in synthetic aperture radar images with a sliding hidden Markov chain model. J. Appl. Remote Sens. **2**(1), 023,526 (2008)
7. Celeux, G., Diebolt, J.: The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comput. Stat. Q. **2**, 73–82 (1985)

8. Delmas, J.P.: An equivalence of the EM and ICE algorithms for exponential family. In: IEEE Trans. Signal Process. **45**(10), 2613–2615 (1997)
9. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Ser B **39**, 1–38 (1977)
10. Derrode, S., Pieczynski, W.: Signal and image segmentation using pairwise Markov chains. IEEE Trans. Signal Process. **52**(9), 2477–2489 (2004)
11. Devijver, P.: Baum's forward-backward algorithm revisited. Pattern Recogn. Lett. **3**, 369–373 (1985)
12. Efron, B.: Bootstrap method: another look at the Jackknife. Ann. Stat. **7**, 1–26 (1979)
13. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, CRC Press, New York (1993)
14. Ephraim, Y.: Hidden markov processes. IEEE Trans. Inf. Theory **48**(6), 1518–1569 (2002)
15. Fjørtoft, R., Delignon, Y., Pieczynski, W., Sigelle, M., Tupin, F.: Unsupervised classification of radar images using hidden Markov chains and hidden Markov random fields. IEEE Trans. Geosci. Remote Sens. **41**(3), 675–686 (2003)
16. Ghahramani, Z., Jordan, M.: Factorial hidden Markov models. Mach. Learn. **29**, 245–273 (1997)
17. Giordana, N., Pieczynski, W.: Estimation of generalized multisensor hidden Markov chain and unsupervised image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **19**(5), 465–475 (1997)
18. Lanchantin, P., Lapuyade-Lahorgue, J., Pieczynski, W.: Unsupervised segmentation of triplet Markov chains hidden with long-memory noise. Signal Process. **88**(5), 1134–1151 (2008)
19. Lapuyade-Lahorgue, J., Pieczynski, W.: Unsupervised segmentation of new semi-Markov chains hidden with long dependence noise. Signal Process. **90**(11), 2899–2910 (2010)
20. McLachlan, G., Khrishnan, T.: The EM Algorithm and Extensions. Wiley Series in Probability and Statistics. Wiley, New York (2008)
21. M'Hiri, S., Cammoun, L., Ghorbel, F.: Speeding up HMRF-EM algorithms for fast unsupervised image segmentation by bootstrap resampling: application to the brain tissue segmentation. Signal Process. **87**(11), 2544–2559 (2007)
22. Pieczynski, W.: Statistical image segmentation. Mach. Gr. Vis. **1**(1/2), 261–268 (1992)
23. Pieczynski, W.: Sur la convergence de l'estimation conditionnelle itérative. Comptes Rendus de l'Académie Des Sciences-Mathématique **346**(7/8), 457–460 (2008)
24. Pieczynski, W., Desbouvries, F.: On triplet Markov chains. In: Proceeding of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'05). Brest, France (2005)
25. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)
26. Skarbek, W.: Generalized Hilbert scan in image printing. In: Klette, R., Kropetsh W.G. (Edn.) chap. Theoretical Foundations of Computer Vision. Akademie Verlag, Berlin, Germany (1992)
27. Weber, K., Bengio, S., Bourlard, H.: Increasing speech recognition noise robustness with HMM2. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02), pp. I.929–932. Orlando, Florida, USA (2002)
28. Wilson, A., Bobick, A.: Parametric HMM for gesture recognition. IEEE Trans. Image Process. **8**(9), 884–900 (1999)
29. Zhong, S., Ghosh, J.: HMMs and coupled HMMs for multi-channel EEG classification. In: Proceeding of IEEE International Joint Conference on Neural Networks, pp. 1154–1159. Honolulu, Hawaii, USA (2002)