Combining Monte Carlo and Mean-Field-Like Methods for Inference in Hidden Markov Random Fields

Florence Forbes and Gersende Fort

***Abstract*—Issues involving missing data are typical settings where exact inference is not tractable as soon as nontrivial interactions occur between the missing variables. Approximations are required, and most of them are based either on simulation methods or on deterministic variational methods. While variational methods provide fast and reasonable approximate estimates in many scenarios, simulation methods offer more consideration of important theoretical issues such as accuracy of the approximation and convergence of the algorithms but at a much higher computational cost. In this work, we propose a new class of algorithms that combine the main features and advantages of both simulation and deterministic methods and consider applications to inference in hidden Markov random fields (HMRFs). These algorithms can be viewed as stochastic perturbations of variational expectation maximization (VEM) algorithms, which are not tractable for HMRF. We focus more specifically on one of these perturbations and we prove their (almost sure) convergence to the same limit set as the limit set of VEM. In addition, experiments on synthetic and real-world images show that the algorithm performance is very close and sometimes better than that of other existing simulation-based and variational EM-like algorithms.**

***Index Terms*—Hidden Markov random fields (HMRFs), image segmentation, Markov chain Monte Carlo-based approximations, variational expectation maximization (VEM).**

I. Introduction

MISSING data models are commonly used in various applications including areas as diverse as signal and image processing, genetics and epidemiology. They are very useful in modeling variability and heterogeneity in data and in solving various labeling or clustering issues. Due to the missing data structure, inference, and parameter estimation, tasks in such models often yield procedures that are intractable as soon as nontrivial interactions in the data are taken into account. In most applications, their complexity requires the development of approximations techniques. These techniques are usually based either on deterministic numerical methods such as variational methods (e.g., [1] and [2]) or on simulation methods such as Markov chain Monte Carlo (MCMC) techniques (e.g., [3]). Choosing one or other approach can be advantageous depending on the context and the goal in mind. Inference problems are usually formulated as the computation of a quantity of interest (e.g., a probability distribution) defined as the solution to an optimization problem specified through a cost function and a constraint set over which the optimization takes place. Variational methods then arise as *relaxations*, that is, simplified optimization problems that involve some approximation of the constraint set, the cost function or both. The original issue is replaced by an easier optimization problem and variational methods have been shown to provide fast and reasonable approximate estimates in many scenarios [1]. However, it appears frequently that these approximations are being used on practical problems with little consideration of important issues such as the accuracy of the approximation, convergence of the algorithms and so on. Convergence results exist for the so-called variational expectation maximization (VEM) algorithms (see [4]), but their application is restricted to specific settings which limit the kind of interactions allowed between the missing data to very simple ones. Variants to extend the application domain of algorithms such as VEM have been proposed (see, e.g., [5] and [6] in an image segmentation framework), but they did not succeed in preserving the convergence results. As a matter of fact, in most settings of practical interest, theoretical results regarding accuracy and convergence properties are still missing. Simulation methods appear then as natural candidates to make algorithms tractable for a wider class of problems while providing tools to study their convergence. As an example, the convergence of MCMC based algorithms has been widely studied and a lot of tools are now available that make various convergence results available or at least easy to derive (see, for instance, [7] for a convergence proof of the Monte Carlo EM algorithm of [8] based on Monte Carlo integration procedure with MCMC sampling techniques). In this paper, our aim is to show that combining both type of methods to design new algorithms can greatly improve accuracy and modeling flexibility in missing data settings. The main idea is that algorithms resulting from such a combination will benefit from the good features of both approaches simultaneously. Deterministic schemes are easy to implement and can provide fast estimates while simulation methods often lead to more accurate results with guaranteed convergence. There have been other attempts at combining approximation techniques and simulation methods. The closest in spirit to our approach is that in [9]. The authors introduce a class of MCMC algorithms that use variational approximations as initial proposal distributions and consider an

Manuscript received January 10, 2006; revised August 13, 2006. This work was supported in part by the French Department of Research under the program "ACI Masses de données." The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zoltan Kato.

F. Forbes is with the MISTIS team, INRIA Rhône-Alpes, ZIRST, Montbonnot, 38334 Saint-Ismier Cedex, France (e-mail: florence.forbes@inrialpes.fr).

G. Fort is with the LTCI, CNRS, 75634 Paris Cedex 13, France (e-mail: gfort@tsi.enst.fr).

Digital Object Identifier 10.1109/TIP.2006.891045

application to sigmoidal belief networks. In our work, we use a different approach and different tools. We incorporate MCMC simulation into variational algorithms and focus on a different application. Other attempts in the statistics community include the use of Laplace approximation with simulation techniques [10], the Gibbsian-EM [11], the restoration-maximization algorithm [12], the Monte Carlo approximations by [13], and, more recently, the simulated field algorithm of [6]. However, most of these procedures were not originally designed with this combining idea in mind and no convergence results are available for them. A detailed comparison of some of these algorithms, for the case of hidden binary isotropic Markov chains, can be found in [14].

Image segmentation and hidden Markov random fields (HMRF) estimation is a typical setting where one encounters these tradeoffs between accuracy, convergence guarantees and reasonable computing time. The expectation maximization (EM) algorithm [15], typically used in missing data cases, yields update procedures that do not have a closed form expression and is intractable analytically. Different algorithms have been proposed to overcome this intractability of EM. Among *pure* simulation techniques, a straightforward variant of the Monte Carlo EM algorithm can be used (see Section V) while variational versions of EM are deterministic alternatives. In particular, VEM algorithms have been popular in cases where the E-step of EM is intractable [1]. The most popular class of VEM procedures is certainly the mean-field EM one. The mean field approach consists in computing quantities related to a complex probability distribution, by using a simple tractable model such as the family of independent distributions. However, introducing relaxation in the E-step does not fully answer the question of inference in cases where the M-step remains intractable due to the complex structure of dependence between the hidden variables. It follows that VEM algorithms cannot be directly applied in the HMRF segmentation framework where additional approximations are required in the M-step. Further algorithms have then been designed that propagate the relaxation in the E-step to the M-step. The combination in such a way of the mean field theory and the EM procedures for HMRF is due to [5]. Using ideas from this principle, [6] proposed, in the context of Markovian image segmentation, a class of EM-like algorithms generalizing [5] which show good performance in practice. In this work, we present another way to overcome the intractability of VEM based on the idea of combining deterministic and simulation-based approximations. We start (Section II) from VEM procedures for which convergence properties are well established and introduce simulations in these algorithms. In addition to make the algorithms tractable, we claim that the introduction of a small perturbation at each iteration of VEM, yields algorithms with the same asymptotic behavior as VEM. More specifically, we propose a class of (stochastically) perturbed VEM algorithms where the noise at each iteration is controlled so that it gets negligible, in a sense to be specified, when the number of iterations tends to infinity. We prove our claim by adapting the results of [7] relative to perturbed iterative maps. We propose (Section III) an example of such a stochastic VEM algorithm, the Monte Carlo VEM algorithm (MCVEM) which is tractable in practice and for which we prove convergence results (Section IV). In addition, the algorithm performance is compared (Section V), on synthetic and real-world images, with various other algorithms that are typical of one of the approach separately. For deterministic algorithms, we report the comparison with the mean field algorithm of [6] while for pure MCMC techniques, we consider a simple extension of the MCEM algorithm, the later being intractable in the HMRF setting. As an illustration, we also compare with two other algorithms among the ones that combine simulation and deterministic methods, namely the Gibbsian-EM and the simulated field algorithms, chosen for their flexibility in missing data problems. We observe that the MCVEM algorithm provides the best (or is very close to the best) results for most of our test images. Our algorithm has, thus, many advantages: a) it is tractable in practice, b) we are able to prove convergence results so that the set of its limit points is identified (as being the set of the limit points of VEM), and c) it is efficient when applied to image segmentation. It illustrates how combining deterministic and simulation techniques can result in improved algorithms.

II. Markov Model-Based Image Segmentation and VEM Algorithms

Let S be a finite set of sites with a neighborhood system defined on it. Let $N = |S|$ denote the number of sites. A typical example in image analysis is the 2-D lattice with a first-order neighborhood system: for each site, the neighbors are the four sites surrounding it. A set of sites C is called a clique if it contains sites that are all neighbors. Let V be a finite set with K elements. Each of them will be represented by a binary vector of length K with one component being 1, all others being 0, so that V will be seen as included in $\{0,1\}^{K\times K}$ and its elements denoted by $\{e_1,\ldots,e_K\}$. We define a discrete Markov random field as a collection of discrete random variables, $\mathbf{Z} = \{Z_i, i \in S\}$, defined on S, each Z_i taking values in V, whose joint probability distribution p_Z is a Gibbs distribution given by

$$p_Z(\mathbf{z}) = W^{-1}\exp\left(-H(\mathbf{z})\right) \tag{1}$$

where H is the energy function $H(\mathbf{z}) = \sum_c V_c(\mathbf{z}_c)$; $\mathbf{z}_c$ denotes a realization of the field restricted to clique c and the V_c's are the clique potentials that may depend on parameters, not specified in the notation. $W = \sum_{\mathbf{z}} \exp(-H(\mathbf{z}))$ is the normalizing factor also called the partition function; $\sum_{\mathbf{z}}$ denotes a sum over all possible values of $\mathbf{z}$. The computation of W involves all possible realizations $\mathbf{z}$ of the Markov field. Therefore, it is, in general, exponentially complex, and not computationally feasible. This can be an issue when using these models in situations where an expression of the joint distribution $p_Z(\mathbf{z})$ is required. We will denote by $\mathcal{Z} = V^N$ the set in which $\mathbf{Z}$ takes values and by $\mathcal{D}$ the set of probability distributions on $\mathcal{Z}$.

In this paper, we focus on Markov model-based image segmentation. Image segmentation involves observed variables (e.g., noisy image pixels) and unobserved variables (e.g., unknown class assignments) which have to be recovered. The hidden variables are modeled as a discrete Markov random field, $\mathbf{Z}$, with distribution p_Z as defined in (1) and an energy

function H depending on a parameter $\beta \in \mathcal{B} \subseteq \mathbb{R}$ and henceforth denoted by $H(\mathbf{z};\beta)$. It is assumed that the observations $\mathbf{Y}$ are conditionally independent given the Markov random field $\mathbf{Z}$, with conditional distribution $p_{Y|Z}$ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^{n_\theta}$, where n_θ is the dimension of θ depending on the model under consideration. In the general case, the likelihood of $(\mathbf{Y}, \mathbf{Z})$ $p_{(Y,Z)}$, called the complete likelihood, is given by

$$p_{(Y,Z)}(\mathbf{y},\mathbf{z};\theta,\beta) = p_{Y|Z}(\mathbf{y}|\mathbf{z};\theta)p_Z(\mathbf{z};\beta). \tag{2}$$

Then the conditional likelihood $p_{Z|Y}$ of the hidden variables Z given the observations Y, is given by $p_{Z|Y} = p_{(Y,Z)}/p_Y$ where p_Y is the likelihood of the observations Y (called the *incomplete* likelihood). It is easy to see that, for such a hidden Markov field model, the conditional field $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ is a Markov field as $\mathbf{Z}$ is with energy function $H(\mathbf{z};\beta) - \log p_{Y|Z}(\mathbf{y}|\mathbf{z}\theta)$. Hereafter, we will refer to the Markov fields $\mathbf{Z}$ and $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ as the marginal and the conditional fields.

In image segmentation problems, the question of interest is generally to recover the unknown image $\mathbf{z}$, interpreted as a classification into a finite number K of labels. This classification usually requires values for the vector parameter $\psi = (\theta, \beta)$. If unknown, the parameters are usually estimated in the maximum likelihood sense

$$\hat{\psi} = \text{argmax}_{\psi\in\Psi} \ln p_Y(\mathbf{y};\psi) \tag{3}$$

where $\Psi = \Theta \times \mathcal{B}$ is the parameter space. This optimization is usually solved by the iterative EM procedure [15]. Any iteration may be formally decomposed into two steps: given the current value of the parameter ψ^t, the so-called E-step consists in computing the expectation of the complete log-likelihood knowing the observations $\mathbf{y}$ and the current estimate ψ^t. In the M-step, the parameter is then updated by maximizing this expected complete log-likelihood

$$\psi^{t+1} = \text{argmax}_{\psi\in\Psi} \sum_{\mathbf{z}\in\mathcal{Z}} \ln p_{(Y,Z)}(\mathbf{y},\mathbf{z};\psi) p_{Z|Y}(\mathbf{z}|\mathbf{y};\psi^t). \tag{4}$$

It is known that, under mild regularity conditions, EM converges to the set of the stationary points of the incomplete likelihood $\psi \mapsto p_Y(\mathbf{y};\psi)$ [16]. As discussed in [17] and [18], EM can be viewed as an alternating maximization procedure of a function F defined, for any probability distribution $q \in \mathcal{D}$, by

$$F(q,\psi) = \sum_{\mathbf{z}\in\mathcal{Z}} \ln\left(\frac{p_{(Y,Z)}(\mathbf{y},\mathbf{z};\psi)}{q(\mathbf{z})}\right) q(\mathbf{z}). \tag{5}$$

Starting from the current value $(q^t, \psi^t) \in \mathcal{D} \times \Psi$, set

$$q^{t+1} = \text{argmax}_{q\in\mathcal{D}} F(q,\psi^t) \tag{6}$$

and

$$\begin{aligned}\psi^{t+1} &= \text{argmax}_{\psi\in\Psi} F(q^{t+1},\psi) \\ &= \text{argmax}_{\psi\in\Psi} \sum_{\mathbf{z}\in\mathcal{Z}} \ln p_{(Y,Z)}(\mathbf{y},\mathbf{z};\psi) q^{t+1}(\mathbf{z}).\end{aligned} \tag{7}$$

The first optimization (6) has an explicit solution $q^{t+1} = p_{Z|Y}(\cdot|\mathbf{y};\psi^t)$ so that the optimization in (4) and (7) are equal. Hence, the "marginal" sequence $\{\psi^t\}_t$ of the sequence $\{(q^t,\psi^t)\}_t$ produced by the alternating maximization procedure is an EM path. The maximization (7) can also be understood as the minimization of a Kullback–Leibler divergence, up to some convention on p_Y, thus justifying the name of alternating minimization procedure often found in the literature (e.g., [4] and [17]).

There exist different generalizations of EM when the M-step (4) is intractable; it can be relaxed by requiring just an increase rather than an optimum. This yields Generalized EM (GEM) procedures ([19]; see also [20] for a convergence result).

Unfortunately, EM (or GEM) is not appropriate for solving the optimization problem (3) in HMRF due to the complex structure of the hidden variables Z; the distribution $p_Z(\mathbf{z};\beta)$ is only known up to a multiplicative constant (the partition function) that depends upon the parameter of interest β and the domain $\mathcal{Z}$ is too large so that the E-step is intractable. Alternative approaches were proposed and they can be understood as generalizations of the alternating maximization procedures mentioned above: the optimization (6) is solved over a restricted class of probability distribution $\tilde{\mathcal{D}}$ on $\mathcal{Z}$ and the M-step (7) remains unchanged. This yields the variational EM (VEM) algorithms [1]. VEM can also be introduced as resulting from a relaxation of a convex optimization problem; the objective function $p_Y(\mathbf{y};\cdot)$ is re-written as the ratio of two partition functions and VEM results from the approximation of one of them using the notion of conjugate duality in convex analysis (see [21] and [2] for details).

Reference [4] proved that, under mild regularity conditions, VEM converges to the set $\mathcal{L}$ of the stationary points of the function F in $\tilde{\mathcal{D}}$. Here, again, generalizations of VEM can be defined by requiring an increase rather that an optimum in the M-step (7), thus defining generalized VEM procedures. These relaxation methods are part of the generalized alternating minimization procedures [4]. The most popular form of VEM is the case when $\tilde{\mathcal{D}}$ is the set of the independent probability distributions on $\mathcal{Z}$ so that $q^{t+1}(\mathbf{z})$ is a factorized distribution $\prod_{i\in\mathcal{S}} q_i^{t+1}(z_i)$. Optimizing (6) with regard to $q_i^{t+1}(e_k)$, $i \in S$ and $e_k \in V$ leads to a fixed-point equation

$$\begin{aligned}\forall i \in S, \forall e_k \in V, \quad \ln q_i^{t+1}(e_k) = c_i + \sum_{\mathbf{z}\in\mathcal{Z}} \ln p_{Z|Y}(\mathbf{z}|\mathbf{y};\psi^t) \\ \times \left\{ \delta_{e_k}(z_i) \prod_{j\neq i} q_j^{t+1}(z_j) \right\}\end{aligned} \tag{8}$$

where c_i is the normalizing constant and δ_e denotes the Dirac mass at point e. The Markov property implies that the right-hand side of the equation only involves the probability distributions q_j, for all j in the neighborhood of i. Existence and uniqueness of a solution to (8) are properties that have not yet been fully understood and will not be discussed here. We refer to [22] for a better insight into the properties of the (potentially multiple) solutions of the mean field equations. Such solutions are usually computed iteratively (see [23]–[25] and an erratum [26]). We will discuss in Section V the consequences of the nonunicity

of the solution when running mean-field based procedures for image segmentation.

Despite the relaxation which may make the summation of the VEM E-step explicit for a convenient choice of $\tilde{\mathcal{D}}$ [i.e., the computation of $F(q^{t+1},\psi)$ in (7)], VEM remains intractable for hidden Markov random fields. From (2) and (7), θ and β are updated independently, given q^{t+1}. Under additional commonly used assumptions on $p_{Y|Z}$, θ^{t+1} is computed in closed form (see for example Section V). The issue is the update of β since it requires an explicit expression of the partition function or an explicit expression of some related quantities (its gradient for example).

To overcome this difficulty, different approaches have been proposed. The *Mean Field* and *Simulated Field* algorithms proposed in [6] are alternatives to VEM that propagate the approximation q^{t+1} of $p_{Z|Y}(\mathbf{z}|\mathbf{y},\psi^t)$ to $p_Z(\mathbf{z};\beta)$. The approach we propose here differs in that the approximation method does not lead to a simple valid model but appears as a succession of approximations to overcome successive computational difficulties. We now turn to this new method.

III. Monte Carlo VEM Algorithm

The theoretical contribution of this paper states that introducing noise at each VEM iteration in such a way that this perturbation goes to zero (in a sense to be specified) as the number of iterations increases, yields an algorithm which has the same asymptotic behavior as VEM (see Section IV). This noise is defined in order to make VEM tractable for solving inference in hidden MRF. We propose an example of such procedures: our stochastically perturbed version of VEM consists in approximating the partition function $W(\beta)$ by some Monte Carlo sum. This yields the so-called Monte Carlo VEM (MCVEM) algorithm. Due to the simulation step, MCVEM is a stochastic algorithm. A difficulty, when dealing with random sequences $\{(q^t,\psi^t)\}_t$ is to guarantee the almost-sure boundedness. Under suitable assumptions (see Appendix I-A), the VEM sequence remains compact so that MCVEM sequences are almost surely bounded provided the Monte Carlo approximations are good enough. The stabilization of MCVEM can be done as described in [7] for the stabilization of the Monte Carlo EM. This corresponds to Step iv) below. The step consists in introducing a variable τ^t which counts the number of re-projections from time 0 to time t ($\tau^0=0$) (see comments below).

Let $\mathcal{I}$ be the set of independent probability distributions on $\mathcal{Z}$ and $\mathcal{I}_r$ be the set of independent probability distributions on $\mathcal{Z}$ such that $q\in\mathcal{I}_r$ implies that $\forall e_k\in V$, $\sum_{i=1}^N q_i(e_k)\neq 0$; $\mathcal{I}_r$ contains the independent probability distributions on $\mathcal{Z}$ such that the probability that no pixels are labeled k is zero.

Fix a positive sequence $\{\gamma^t\}_t$ such that $\inf_t\gamma^t>0$, a sequence of nondecreasing positive integers $\{J_t\}_t$ and a sequence of probability distributions $\{\pi^t\}_t$ on $\mathcal{Z}$. Let $\{\mathcal{C}^t\}_t$ be a sequence of compact subsets such that for any $t\geq 0$

$$\mathcal{C}^t\subsetneq\mathcal{C}^{t+1}\qquad \mathcal{I}_r\times\Psi=\bigcup_{t\geq 0}\mathcal{C}_t \tag{9}$$

and set $\tau^0=0$. For the current value (q^t,ψ^t) of the parameter.

i) Update the q-component

$$q^{t+1}\in\operatorname{argmin}_{q\in\mathcal{I}_r}\sum_{\mathbf{z}}\log\left(\frac{q(\mathbf{z})}{p_{Z|Y}(\mathbf{z}|\mathbf{y};\psi^t)}\right)q(\mathbf{z}).$$

ii) Update the θ-component

$$\theta^{t+1}\in\operatorname{argmax}_{\theta\in\Theta}\sum_{\mathbf{z}}\ln p_{Y|Z}(\mathbf{y}|\mathbf{z};\theta)q^{t+1}(\mathbf{z}).$$

iii) Sample a Markov random field $\{Z^{j,t}\}_{1\leq j\leq J_t}$ with invariant distribution $\pi^t(\mathbf{z})$. Set

$$\beta^{t+1}\in\operatorname{argmax}_{\beta\in\mathcal{B}^t}-\left\{\sum_{\mathbf{z}\in\mathcal{Z}}H(\mathbf{z};\beta)q^{t+1}(\mathbf{z})+\ln\tilde{W}^{J_t,\pi^t}(\beta)\right\} \tag{10}$$

where

$$\tilde{W}^{J_t,\pi^t}(\beta)=\frac{1}{J_t}\sum_{j=1}^{J_t}\exp\left(-H(\mathbf{z}^{j,t};\beta)-\ln\pi^t(\mathbf{z}^{j,t})\right)$$

and $\mathcal{B}^t=\{\beta\in\mathcal{B},|\beta-\beta^t|\leq\gamma^t\}$.

iv) If $(q^{t+1},\psi^{t+1})\notin\mathcal{C}^{\tau_t}$, re-initialize the parameter by setting $q^{t+1}=q^0$ and $\psi^{t+1}=\psi^0$; and increment the counter $\tau^{t+1}=\tau^t+1$. Otherwise, set $\tau^{t+1}=\tau^t$.

In practice, Step i) is implemented directly by iteratively solving a nonlinear system given by the mean field (8) (see [25]).

Step iii) looks like the algorithm proposed in [27] for maximum likelihood parameter estimation of exponential families, except that, in [27], it is assumed that the samples are independent and identically distributed. In (10), when the maximum is unique, the β-update follows the MCMCML algorithm proposed in [28] for the estimation of fully observed Markov Random Field prior parameters. We will describe in Section V models for which the maximum is unique.

In MCVEM, the idea is that the partition function is approximated by a Monte Carlo sampling from some distribution π^t, thus using an importance sampling estimator (or possibly a self-normalized importance sampling estimator [29]). If the sampler is good enough so that a law of large numbers holds, $\lim_{J\to\infty}\tilde{W}^{J,\pi^t}(\beta)=W(\beta)$, and one can expect that by choosing J large enough, $\tilde{W}^{J,\pi^t}(\beta)$ provides a good approximation of $W(\beta)$. As discussed in [30], the best choice for approximating $W(\beta)$ is $\pi=p_Z(\cdot;\beta)$. This is useless for our purposes since we want a good approximation of $W(\beta)$ whatever β, for a given distribution π. Nevertheless, by choosing $\pi^t=p_Z(\cdot;\beta^t)$, it can be expected that for some J sufficiently large, $\tilde{W}^{J,\pi^t}(\beta)$ is a good approximation of the partition function $W(\beta)$ in a neighborhood of β^t. This explains the local optimization in (10) and the introduction of the domain $\mathcal{B}^t$. $\{\gamma^t\}_t$ is a deterministic sequence uniformly bounded away from zero. One could be interested in choosing γ^t as a function of the Hessian of the quantity optimized in (10); in that case, $\{\gamma^t\}_t$ is a random sequence and the study of the asymptotic behavior of MCVEM is slightly more complex. This extension is left to the interested reader.

In practice, we choose $\pi^t = p_Z(\cdot;\beta^t)$ and sample the Markov random field by using Markov chain Monte Carlo samplers. Given a distribution π known up to a scaling factor, MCMC samplers consist in sampling a Markov chain from a transition kernel defined such that π is its unique stationary distribution. Under suitable conditions on π and on the transition kernel, a strong law of large numbers holds for a large family of functions [3]. Among the most usual MCMC samplers when π is a Gibbs distribution, are the Gibbs sampler [31], the Hastings–Metropolis sampler [32], or the Swendsen–Wang sampler [33]. Observe that $\tilde{W}^{J,\pi^t}(\beta)$ can be known up to a multiplicative constant independent of β: this allows the choice $\pi^t = p_Z(\cdot;\beta^t)$ which is known up to the partition function $W(\beta^t)$. Hereafter, we will simply write $\tilde{W}^{J,\beta^t}$ as a shorthand notation for $\tilde{W}^{J,p_Z(\cdot;\beta^t)}$. We have found this simple method for estimating $W(\beta)$ easy to use and very satisfying in our experiments that we chose as typical segmentation problems (see Section V). However, we are aware of possible limitations of such MCMC samplers. In practice our numerical results could certainly be further improved by using more sophisticated methods. A full analysis of the problem of estimating normalizing constants has been given in [34]. They discussed several methods that are more sophisticated but also more cumbersome. In this paper, our focus is mainly on convergence results and on showing that it is advantageous to combine variational and MCMC methods. We did not investigate further the possibility of using better samplers.

Step iv) can be understood as a random re-initialization of the algorithm. At iteration $t+1$, the candidate (q^{t+1},ψ^{t+1}) has to be in the compact set $\mathcal{C}^{\tau^t}$; otherwise, the sequence $\{(q^t,\psi^t)\}_t$ is re-initialized and the counter incremented, so that for the next step, the parameter sequence is allowed to be in a larger compact set. Observe that each time the counter is incremented, the sequences $\{\gamma^t\}_t$ and $\{J_t\}_t$ are not re-initialized: since $\{J_t\}_t$ is nondecreasing, it follows a larger number of simulations at each iteration. We can, thus, expect a better Monte Carlo approximation, thus explaining that the MCVEM sequences inherit the boundedness property of the VEM sequences.

We refer to Section V for an illustration of a suitable choice of the different implementation parameters (such as $\{J_t\}_t$, $\{\gamma^t\}_t,\cdots$).

IV. Convergence Theorems for Stochastically Perturbed VEM Algorithms

We provide sufficient conditions on the model and on the Monte Carlo approximations ensuring that MCVEM and VEM have the same asymptotic behavior: the set of the limit points of MCVEM is the set of the limit points of VEM. The conditions and the proofs of our claims are quite technical. For clarity, we postpone in Appendix I the assumptions and a rigorous statement of our results, while the detailed proofs are given in [35]. In this section, we comment the assumptions and the theoretical results. We also give pointers to the key elements that need to be considered for the proof of convergence to still hold when deriving algorithms that can be read as stochastically perturbed VEM algorithms. We show in Appendix II how the assumptions are satisfied when MCVEM is used for a nontrivial application.

The key idea of the proof of convergence of MCVEM is that this algorithm is a stochastically perturbed VEM algorithm: the perturbations come from the Monte Carlo approximations. Under a convenient choice of MCMC samplers, these perturbations vanish as the number of iterations goes to infinity since the number of simulations per iteration $\{J_t\}_t$ increases. Convergence of VEM relies on the existence of a Lyapunov function (see, e.g., [36] for a definition), namely $L = \exp(F)$ where F is given by (5). This remains true for the generalized VEM algorithm obtained by replacing, in the update of the β component, the global optimization by a local one on the domain $\mathcal{B}^t$. Furthermore, this generalized VEM and VEM have the same limit points. Unfortunately, due to the introduction of a perturbation at every iteration of the (generalized) VEM algorithm, the function L is not a Lyapunov function for MCVEM. Nevertheless, since the perturbation vanishes, we expect MCVEM to inherit the asymptotic behavior of the generalized VEM. The first set of conditions (see **A1**–**A3** in Appendix I) is relative to the model and ensures that the function L is a Lyapunov function for the generalized VEM algorithm. This implies that $L(q^{t+1},\psi^{t+1}) \geq L(q^t,\psi^t)$ where $\{(q^t,\psi^t)\}_t$ is a generalized VEM sequence. Since $\sup_{\{(q,\psi),L(q,\psi)\geq L(q^0,\psi^0)\}} L$ is assumed to be compact, the sequence $\{L(q^t,\psi^t)\}_t$ converges to some point L^* of the form $L^* = L(q^*,\psi^*)$ where (q^*,ψ^*) is in $\mathcal{L}$, the set of fixed points of the VEM algorithm

$$\begin{aligned}\mathcal{L} = \{&(q^*,\psi^*) \in \mathcal{I}_r \times \Psi,\\ &q^* \in \operatorname{argmin}_{q\in\mathcal{I}_r} \sum_{\mathbf{z}} \log\left(\frac{q(\mathbf{z})}{p_{Z|Y}(\mathbf{z}|\mathbf{y};\psi^*)}\right) q(\mathbf{z})\\ &\text{and}\quad \psi^* \in \operatorname{argmax}_{\psi\in\Psi} F(q^*,\psi)\}\,. \end{aligned} \tag{11}$$

The second set of conditions (see **A5** and **A6** in Appendix I) is relative to the Monte Carlo approximations. We require the MCMC samplers to be such that the L^r-errors when approximating exact expectations by Monte Carlo sums with J terms, decrease to zero at rate $J^{-r/2}$ (for some $r \geq 2$, see **A5**). Furthermore, the number of simulations J_t has to increase all the more so as r is small (see **A6**). Condition **A4** is to quantify the increase of the Lyapunov function after one iteration of VEM, when started outside any open neighborhood of the limit set $\mathcal{L}$. Under these conditions, we can apply a result provided in [7], which is the central tool of our proof. Let $\mathcal{K}$ in $\mathcal{I}_r \times \Psi$ be a compact set, and starting from (q^t,ψ^t), denote by (q^{t+1},ψ^{t+1}) the new parameter after one iteration of MCVEM (respectively, $(q^{t+1},\bar{\psi}^{t+1})$, after one iteration of the generalized VEM algorithm). The result states that if

$$\lim_t \left|L(q^{t+1},\psi^{t+1}) - L(q^{t+1},\bar{\psi}^{t+1})\right| \mathbb{1}_{(q^t,\psi^t)\in\mathcal{K}} = 0 \tag{12}$$

almost surely when the perturbation is stochastic, *(a)* the number of re-projections in Step iv) is almost surely finite and *(b)*, MCVEM and VEM have the same asymptotic behavior (in some sense). Condition (12) means that the perturbation vanishes along the compact path, when measured in terms of the error induced on the Lyapunov function. The main step of our proof consists in proving that under the stated assumptions on the Monte Carlo approximations, this condition is satisfied.

We provide two convergence results. The first one (Theorem 1 in Appendix I) shows that the generalized VEM algorithm and the VEM algorithm have the same limit points. The second one (Theorem 2 in Appendix I), which is the original theoretical contribution of this paper, states that for almost all trajectories of MCVEM, *(a)* the number of re-projections is finite and the path remains in a compact set and *(b)*, the sequence $\{L(q^t, \psi^t)\}_t$ converges to a subset of $\mathcal{L}$. Combined with Theorem 1, it also implies that under a suitable condition on the interior of the set $L(\mathcal{L})$, the MCVEM algorithm, the generalized VEM algorithm and the VEM algorithm have the same asymptotic behavior. In all cases, the sequence $\{L(q^t, \psi^t)\}_t$ converges to $L^* = L(q^*, \psi^*)$ for some (q^*, ψ^*) in the solution set $\mathcal{L}$, and the paths $\{(q^t, \psi^t)\}_t$ produced by the MCVEM procedure, the generalized VEM one or the VEM one, converge to some subset of $\mathcal{L}$.

As already mentioned (see also the comments in Appendix I), the convergence claims for MCVEM are based on an extension of the results in [7]. These results could be applied to address the convergence of any (stochastic) perturbation of an iterative algorithm that admits a Lyapunov function. More specifically, we denote by T the point-to-set map associated to the iterative procedure having a Lyapunov function L relative to a set $\mathcal{L}$. Let $\{u^t\}_t$ be the (possibly stochastic) sequence produced by the perturbed algorithm. If $(T, \{u^t\}_t)$ satisfies the three conditions of [7, Proposition 10], then *(a)* the sequence $\{u^t\}_t$ is compact (say, in $\mathcal{K}$), *(b)* the sequence $\{L(u^t)\}_t$ converges to a connected component of $L(\mathcal{K} \cap \mathcal{L})$, and *(c)* if this set has an empty interior, the sequence $\{u^t\}_t$ converges to a subset of $\mathcal{L}$. The first two conditions of [7, Proposition 10] are relative to the Lyapunov function. The third condition requires $\{u^t\}_t$ to be infinitely often in a compact set and, for any compact set $\mathcal{K}$, $|L(u^{t+1}) - L(w)| 1_{u^t \in \mathcal{K}}$ tends to zero for some $w \in T(u^t)$. These convergences are almost sure when $\{u^t\}_t$ is a random sequence. When applied to MCVEM, this corresponds, respectively, to **A1**–**A4** in Appendix I, to the recurrence condition implied by Step iv) and to (12). As a conclusion, we point out that the main assumption to address the convergence of an iterative algorithm, understood as a perturbation of a procedure having a Lyapunov function, is condition (12).

V. Application to Image Segmentation

In this section, we turn more specifically to the applications. MATLAB codes for MCVEM are available on the web page of the authors. We consider simple models and use a K-color Potts model as the distribution of the hidden fields. Each z_i takes one of K states, which can represent K different class assignments. Each of them is represented by a binary vector of length K with one component being 1, all others being 0. The distribution of a K-color Potts model is defined by

$$p_Z(\mathbf{z}; \beta) = W(\beta)^{-1} \exp\left(\beta \sum_{i \sim j} z_i^t z_j\right)$$

where the notation $i \sim j$ represents all couples of sites (i, j) which are neighbors. Parameter β is a spatial parameter that controls the strength of the interaction between neighboring sites. In a segmentation framework, the Potts model acts as a regularizing (smoothing) term. The lower β, the weaker the regularization. The factorized conditional distribution $p_{Y|Z}(\mathbf{y}|\mathbf{z}; \theta)$ is of the form $p_{Y|Z}(\mathbf{y}|\mathbf{z}; \theta) = \prod_{i \in S} f_i(y_i|z_i; \theta)$ where f_i is a univariate Gaussian distribution: if z_i is in class k, f_i is the Gaussian distribution with parameters μ_k and σ_k, μ_k and σ_k being the mean and the standard deviation. The parameter to be estimated is then (β, θ) with $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$.

For the simulation step of *MCVEM*, we use the Gibbs sampler. For such models and this sampler, we show in Appendix II that the various assumptions are satisfied so that the previous convergence results apply : any MCVEM path converges, and the set of the limiting values is the set of the limiting values of VEM. Furthermore, assumption A5 is actually satisfied with any initial distribution λ and any $r \geq 2$. Hence, there are no restrictions for the initialization of the Markov chains at each iteration, and suitable choices of J_t are any polynomially increasing sequences.

We compare MCVEM to different algorithms when applied to parameter estimations and image segmentations. We first run an EM procedure (hereafter called *ind-EM*), assuming that missing data are independent, in order to illustrate the gain in taking into account the spatial information. The following other procedures are based on models assuming dependencies. As a typical simulation method, we run a kind of Monte Carlo EM (hereafter *MC2-EM*) where two Monte Carlo approximations are introduced at each iteration. The first one corresponds to the MCEM algorithm [8] and the second one makes the M-step tractable by approximating the partition function $W(\beta)$ as in MCVEM. By combining the convergence results of MCVEM (Section IV) and of MCEM [7], it can be established that *MC2-EM* converges almost-surely to the stationary points of the incomplete log-likelihood $\ln p_Y(\mathbf{y}; \psi)$ and due to its stochastic nature, converges to a (local) maximum [7]. *MC2-EM* has a much higher computational cost but it provides reference solutions to assess the proximity of the MCVEM limiting values to the maxima of the incomplete log-likelihood. We then compare to the *Mean Field* algorithm of [6], as a typical deterministic variational algorithms. Finally, we run two other algorithms designed to overcome the intractability of EM in hidden MRF, the Gibbsian-EM [11] that combines Monte Carlo techniques and pseudo-likelihood approximation and the *Simulated Field* of [6]. The latter two can also be seen as combinations of simulations and deterministic approximations but are not part of the novel strategy we propose. No convergence results are available for them.

In addition to parameter estimation, the way the segmentation task is carried out in the different procedures can vary. For *MCVEM*, *Mean Field*, and *Simulated Field* algorithms, images are restored by using the *maximum a posteriori* (MAP) principle based on the factorized distribution q^{t+1} that approximates the conditional distribution $p_{Z|Y}(\cdot|\mathbf{y}; \psi^t)$. *Gibbsian-EM* and *MC2-EM* both generate realizations of the conditional field and the image reconstruction is performed using the maximizer of the posterior marginal (MPM) decision rule [37]. Note that, for the first three algorithms, the MAP and MPM rules coincide when applied to q^{t+1} since q^{t+1} is a factorized distribution.

For the Potts models, we assumed a first-order neighborhood (four neighbors per pixel). For the stochastic algorithms (i.e., all but *ind-EM* and *Mean Field*), we report the mean values of

the estimates along the random path, where the mean is over the iterations after the burn-in period. Regarding the segmentation results, the error rate (i.e., the proportion of misclassified pixels) corresponds to the mean error rate computed after the burn-in period.

A. Practical Implementation of MCVEM

Prior to any performance comparison, we discuss implementation details of MCVEM such as the initialization of the Markov chain at each iteration, the choice of the simulation scheme $\{J_t\}_t$ and of the sequence $\{\gamma^t\}_t$. As an illustration, the algorithm is run on a 133×142 noise-corrupted two-color image (shown in Fig. 3). We used Gaussian densities with class-dependent means and standard deviation so that the true noise parameters (μ_1, σ_1) and (μ_2, σ_2) are known. We then consider the parameters estimates as a function of the number of iterations when the Markov chain is, at each iteration, initialized at the same point or at the last sample drawn at the previous iteration. The considered J_t is $J_t \sim (2t)^{1.2}$. In the two cases, the results, in terms of parameter estimation and segmentation, are similar but the convergence when using the first strategy is very slow. The same observation holds for other choices of J_t so that in what follows, only the second strategy will be kept.

We then consider different schemes for J_t, namely $J_t \sim (2t)^{1.01}$, $J_t \sim (2t)^{1.3}$, and $J_t \sim (2t)^{1.5}$. All schemes result in a convergence to the same value of β. The value of β in average is not sensitive to the scheme but its variation is all the smaller as the rate is higher (a phenomenon already mentioned in [7]). For the image in Fig. 3, the mean values, computed when the curves stabilize, are equal to 0.93 while the standard deviations are respectively 0.0017, 0.0009, and 0.0006. Similar behavior was observed on other images suggesting not surprisingly that limiting the number of simulations has a cost in that it produces paths with higher variations. In the following developments, we will consider $J_t \sim (2t)^{1.3}$.

We then study the robustness of *MCVEM* to the choice of the starting parameter values (θ^0, β^0) and to the choice of the sequence $\{\gamma^t\}_t$. We consider a constant $\gamma^t = \gamma$ over the iterations. We consider in turn three cases ($\theta^0 = \theta_{km}, \beta^0 = 1, \gamma = 0.005$), ($\theta^0 = \theta_{km}$, $\beta^0 = 5$, $\gamma = 0.05$), and ($\theta^0 = \theta_{thr}$, $\beta^0 = 5$, $\gamma = 0.05$), where θ_{km} and θ_{thr} denote respectively the empirical means and standard deviation corresponding to a k-means classification (displayed in Fig. 3) and those corresponding to a two-color classification obtained by simple thresholding of the image pixels values. The path $\{\beta^t\}_t$ of successive estimations of β is plotted in Fig. 1. We observe that the estimation of $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$ is well performed whatever the algorithm. The plots show that the limiting behavior of *MCVEM* (dotted lines) does not depend on γ, at least when γ is small enough. For large values of γ (say $\gamma = 0.1$), the sequence $\{\beta^t\}_t$ may oscillate for a long time between two values of the form β and $\beta + \gamma$. This illustrates the fact that $\tilde{W}^{J_t, \beta^t}$ can be considered as a reasonable approximation of $W(\beta)$ in a neighborhood of β^t, and justifies the introduction of a local optimization domain $\mathcal{B}^t$ in the update of β. This local optimization explains the linear path of *MCVEM* in the first iterations. These plots illustrate that *MCVEM* is very robust to initialization.

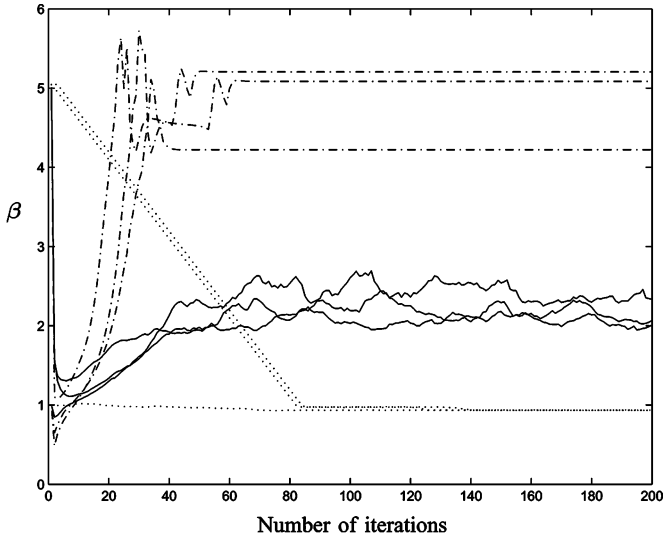

Fig. 1. Logo image: β trajectory versus the number of iterations for different parameter starting values, with (dashed-dotted line) *Mean Field*, (solid line) *Simulated Field*, and (dotted line) *MCVEM*.

For comparison with the two other variational methods we consider, we also run the *Mean Field* and *Simulated Field* algorithms and show the results on the same Fig. 1. It appears that the starting value is crucial for the limiting behavior of *Mean Field* (dashed-dotted lines). On some other synthetic images (not shown here), *Mean Field* actually fails to converge even with reasonable initializations such as those provided by running a *k-means* algorithm. The trajectories of *Simulated Field* (solid lines) do not converge to some fixed limiting value but the behavior of the different trajectories is similar. The same kind of phenomenon was already pointed out in [14] for the *Restoration-Maximization* algorithm close in spirit to the *Simulated Field* algorithm. We believe that convergence of the *Simulated Field* algorithm has to be understood in a different way. An approach similar to what is done for the so-called stochastic EM algorithm is more appropriate (see [38] and [39]). The sequence $\{(q^t, \psi^t)\}_t$ is a realization of a Markov chain and the asymptotic behavior of this sequence is related to the ergodic behavior of this Markov chain. Hence, averages of the parameters should converge and this suggests to replace the current implementation of *Simulated Field* algorithm by an averaging procedure [40]. However, such extensions are beyond the scope of this paper and we run the algorithm as described in [6]. Despite the variations in the estimation of the spatial parameter β, the corresponding segmentations are quite stable: the mean error rate is in the range (2.86%, 2.92%) for *MCVEM* [respectively, (2.82%, 3.10%) for *Mean Field* and (3.42%, 3.65%) for *Simulated Field*].

We finally discuss how the possible nonunicity of q^{t+1}, the mean-field approximation of the conditional field $p_{Z|Y}(\cdot|\mathbf{y}; \psi^t)$, may affect the resulting image segmentations. To that goal, we compute the mean error rates for the segmented images when β is assumed to be known but $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$ is unknown. In Fig. 2, we plot these mean error rates versus β for two different starting points corresponding respectively to a k-means and a thresholding classification as above. These plots

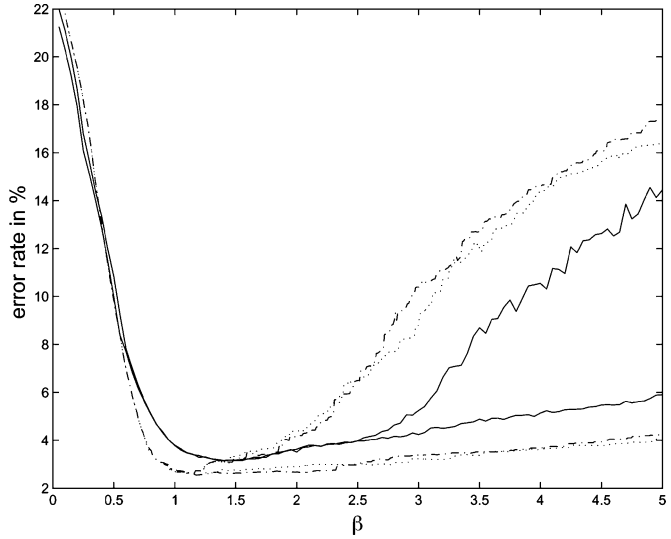

Fig. 2. Logo image : Error rate versus β obtained by (dashed-dotted line) *Mean Field*, (solid line) *Simulated Field*, and (dotted line) *MCVEM*, when the segmentation algorithm is started from two different initial classifications.

show that for large values of β, the segmentation is greatly dependent of the initial segmentation. In addition, the curves give an idea of the β value that corresponds to the minimum error rate. For *MCVEM* and *Simulated Field*, this naive computation is not far from the estimates obtained by running the full algorithms when all the parameters (β, θ) are unknown (these estimations are reported in Table IV). *MCVEM* converges to a limiting value and *Simulated Field* fluctuates around a mean value such that the segmentation is not affected by the nonunicity of q^{t+1}. This is not the case for *Mean Field*, thus showing that the *Mean Field* segmentation may depend on the implementation of the algorithm.

B. Synthetic and Real Images

We now compare in more detail the algorithms' performance when applied to parameter estimation and image segmentation. We report the estimations of β and $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$ and the mean segmentation error rates when a ground truth is available. For comparison, we also indicate in column *ref.* the error rates when the parameters are not estimated but fixed to their true values if known. When given, the corresponding segmentations are computed after the same fixed number of iterations (200) for each iterative algorithm. Three types of test images are presented. Comments on the results are postponed after the description of all experiments.

The algorithms are first tested on images simulated from hidden Potts models for which the true parameters β and θ are known. We created 100×100 images by simulating 2-D K-color Potts models for $K = 2$, 3, 4 and different values of β (lower than the critical value $\beta_c = \ln(1 + \sqrt{K})$), and then adding a Gaussian noise. For each set of parameters we investigate, 20 realizations of each corresponding Potts model are simulated. We then run the different algorithms on these 20 simulations. The results are reported in Tables I–III. The values reported are the mean and standard deviation values over the 20 runs. We note that the estimation of the parameter θ is always satisfying and only the results on β are reported.

TABLE I
PARAMETER ESTIMATES AND ERROR RATES FOR THE HIDDEN TWO-COLOR POTTS MODEL WITH $\beta = 0.78$ (FIRST-ORDER NEIGHBORHOOD). THE RESULTS ARE MEAN VALUES OVER 20 RUNS; THE STANDARD DEVIATIONS ARE ALSO REPORTED IN PARENTHESIS

algorithm	β	error rate	ref.
true values	0.78	-	-
ind-EM	-	15.91 (0.33)	15.85 (0.26)
Mean Field	0.94 (0.0283)	10.28 (0.49)	9.77 (0.42)
Simulated Field	0.78 (0.0224)	10.96 (0.43)	11.04 (0.48)
MCVEM	0.73 (0.0177)	9.87 (0.42)	9.77 (0.42)
MC2-EM	0.77 (0.0144)	9.81 (0.39)	9.81 (0.39)
Gibbsian-EM	0.77 (0.0223)	9.79 (0.40)	9.81 (0.39)

TABLE II
ESTIMATES OF β AND ERROR RATES FOR THE HIDDEN THREE-COLOR POTTS MODEL WITH $\beta = 0.9$ (FIRST-ORDER NEIGHBORHOOD). THE RESULTS ARE MEAN VALUES OVER 20 RUNS; THE STANDARD DEVIATIONS ARE ALSO REPORTED IN PARENTHESIS

algorithm	β	error rate	ref.
true values	0.90	-	-
ind-EM	-	21.31 (0.60)	21.14 (0.50)
Mean Field	1.03 (0.0245)	14.03 (0.60)	13.78 (0.59)
Simulated Field	0.90 (0.0245)	15.67 (0.56)	15.69 (0.64)
MCVEM	0.85 (0.0189)	14.02 (0.59)	13.78 (0.59)
MC2-EM	0.89 (0.0136)	13.77 (0.53)	13.79 (0.54)
Gibbsian-EM	0.89 (0.0223)	13.77 (0.53)	13.79 (0.54)

TABLE III
ESTIMATES OF β AND ERROR RATES FOR THE FOUR-COLOR POTTS MODEL WITH $\beta = 1$ (FIRST-ORDER NEIGHBORHOOD). THE RESULTS ARE MEAN VALUES OVER 20 RUNS; THE STANDARD DEVIATIONS ARE ALSO REPORTED IN PARENTHESIS

algorithm	β	error rate	ref.
true values	1	-	-
ind-EM	-	24.23 (0.54)	23.87 (0.45)
Mean Field	1.05 (0.0195)	18.32 (0.51)	18.38 (0.45)
Simulated Field	0.90 (0.0164)	20.73 (0.55)	20..82 (0.48)
MCVEM	0.81 (0.0117)	18.66 (0.50)	18.38 (0.45)
MC2-EM	0.89 (0.0107)	18.15 (0.49)	18.24 (0.47)
Gibbsian-EM	0.89 (0.0167)	18.14 (0.50)	18.24 (0.47)

The following test images are noise-corrupted images corresponding to known values of K. These images before degradation are not realizations from a known Markov field model. The first image is the logo image described in V-A and shown in Fig. 3. The other example is a 128×128 image obtained by adding some Gaussian noise to the four-color top left image of Fig. 4. The noise parameters are given by $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, 4\}$ with $\mu_k = k$ and $\sigma_k = 0.5$ for $k = 1, \ldots, 4$. The results are reported in Tables IV and V. The corresponding segmentations are shown in Figs. 3 and 4.

We finally run the algorithms on real images for which a true value of K does not exist (in real life, it is usually part of the problem to assess its value) but for which intuition or expert knowledge could give an indication of what would be a reasonable value. As an illustration, the top left image in Fig. 5 is a 76 $\times$ 91 PET image of a dog lung (see [41] for more details on its nature and origin) and the top left image in Fig. 6 is a 256 $\times$ 256

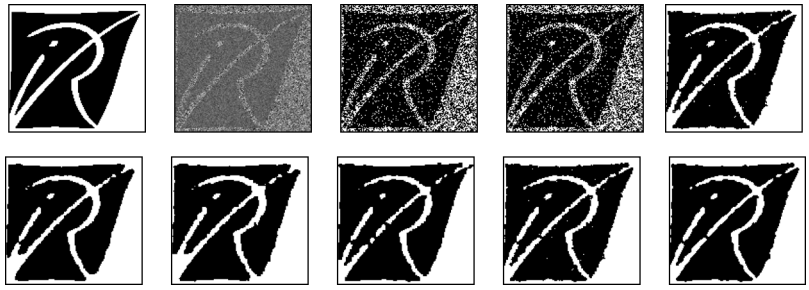

Fig. 3. Logo image: (top, from left to right) original image, noise-corrupted image, initial segmentation using k-means, ind-EM, MC2-EM; (bottom, from left to right) Gibbsian-EM, Simulated Field, Mean Field, MCVEM, MCVEM + Median Filter.

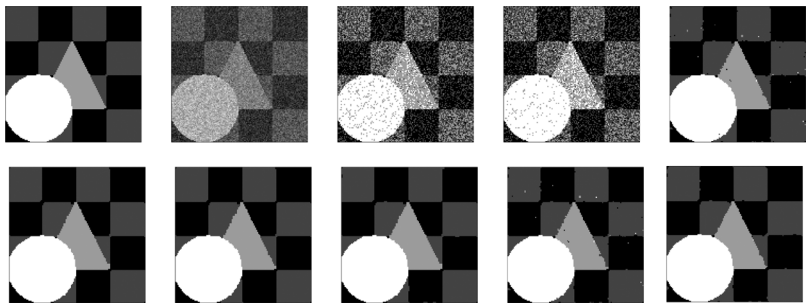

Fig. 4. Triangle image: (top, from left to right) original image, noise-corrupted image, initial segmentation using k-means, ind-EM, MC2-EM; (bottom, from left to right) Gibbsian-EM, Simulated Field, Mean Field, MCVEM, MCVEM + Median Filter.

TABLE IV
PARAMETER ESTIMATES AND ERROR RATES FOR THE DEGRADED TWO-COLOR LOGO IMAGE

algorithm	β	μ_1	μ_2	σ_1	σ_2	error rate
true values	-	51	255	130	300	-
ind-EM	-	52	255	128	304	22.69
Mean Field	4.22	53	260	130	306	3.10
Simulated Field	2.15	52	250	128	302	3.42
MCVEM	0.93	50	262	125	305	2.89
MC2-EM	0.91	50	261	125	305	2.89
Gibbsian-EM	1.82	52	251	128	303	2.92

satellite image. They have been chosen because they correspond to rather different application domains and because nonexpert users can easily assess the quality of their segmentations.

For the dog lung image, the aim is to distinguish the lung from the rest of the image in order to measure the heterogeneity of the tissue in the region of interest. Only pixels in this delimited area are then considered to compute a heterogeneity measure, such as a coefficient of variation. The interpretation of the image suggests that three-color segmentations are reasonable. The image is constructed based on radioactive emissions from gas in the lung. Ideally, the background should correspond to one color and two other colors should account for the high gas density in the interior of the lung and the somewhat lower gas density around the periphery. The resulting segmentations are shown in Fig. 5.

Fig. 6 is a SPOT satellite image representing part of the Aquitaine region in France. It contains large homogeneous regions (large fields, woods), precise contours (rivers, roads) and more heterogeneous areas (houses, small fields) or textured parts. Whether relevant segmentations should focus on contours or regions may depend on the application in mind.

All tables show that *ind-EM* differs from the other algorithms: the estimates are somewhat poor (see, e.g., Table V) and the error rates are much higher. The gain in taking into account spatial dependencies clearly appears.

We observe that the estimation of the means and standard deviations $\{(\mu_k, \sigma_k), k = 1, \dots K\}$ is an easy task in the sense that all algorithms (except *ind-EM*) have similar good performances. We then focus our comments on the estimation of the spatial parameter β which is more critical. When the true value is lower than the critical value β_c, *MCVEM* seems to underestimate β (see Tables I–III). More generally, *MCVEM* provides the lowest estimates, while *Mean Field* provides the highest ones. The *Mean Field* algorithm systematically overestimates β. It is quite difficult to determine which approach is the best, since the value of the spatial parameter β acts upon the image segmentation. Nevertheless, the results of *MC2-EM* which converges to the (local) maxima of the incomplete log-likelihood $\ln p_Y(\mathbf{y}; \psi)$ can be taken as reference values. It appears that *MCVEM* and *MC2-EM* are very close (see Tables IV and V) while *Mean Field* and *Simulated Field* and *Gibbsian-EM* are of a different kind. For the β estimation, *Simulated Field* is close to *Gibbsian-EM* while *Mean Field* is the most atypical. Despite *Simulated Field* and *Gibbsian-EM* rely on different tools (mean-field based variational technique on one hand, pseudo-likelihood approximation on the other hand), they are numerically close. We believe that, due to the ergodicity of the discrete-valued Markov chain which admits the conditional field $p_{Z|Y}$ as invariant distribution, they have indeed very similar asymptotic behaviors.

In terms of segmentation results, *MCVEM* leads to very satisfying error rates: for the hidden Potts images, the error rates

TABLE V
PARAMETER ESTIMATES AND ERROR RATES FOR THE DEGRADED FOUR-COLOR IMAGE

algorithm	β	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	error rate
true values	-	1	2	3	4	0.5	0.5	0.5	0.5	-
ind-EM	-	0.85	1.69	2.54	3.93	0.44	0.42	0.46	0.53	29.4
Mean Field	3.99	0.99	2.00	2.98	4.01	0.49	0.50	0.48	0.50	0.44
Simulated Field	3.46	1.00	2.00	2.97	4.01	0.49	0.50	0.48	0.50	0.40
MCVEM	1.27	0.99	2.00	2.98	4.01	0.48	0.48	0.46	0.50	0.80
MC2-EM	1.26	0.99	2.00	2.97	4.01	0.48	0.48	0.46	0.49	0.81
Gibbsian-EM	3.01	1.00	2.00	2.98	4.00	0.49	0.50	0.48	0.50	0.31

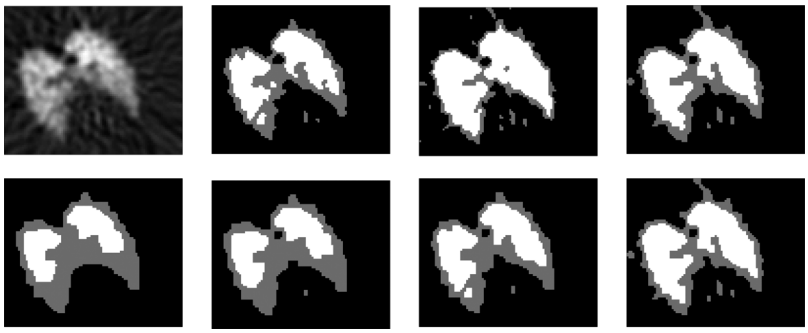

Fig. 5. PET image of a dog lung: (top, from left to right) original image, initial segmentation, *ind-EM*, *MC2-EM*; (bottom, from left to right) *Gibbsian-EM*, *Simulated Field*, *Mean Field*, *MCVEM*.

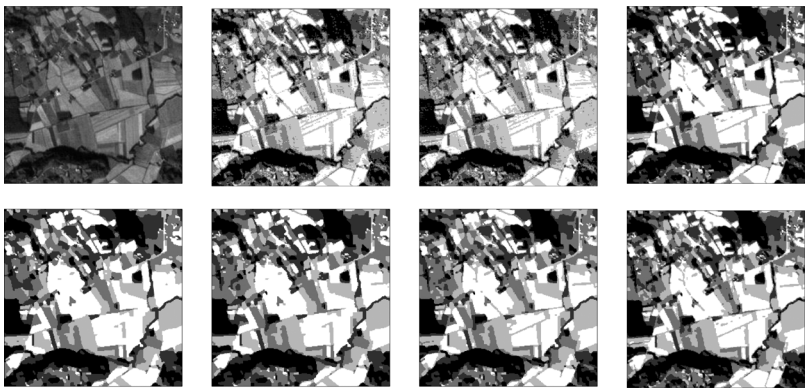

Fig. 6. Satellite image: (top, from left to right) original image, initial segmentation, *ind-EM*, *MC2-EM*; (bottom, from left to right) *Gibbsian-EM*, *Simulated Field*, *Mean Field*, *MCVEM*.

are close to the minimal error rates (achieved with *MC2-EM* and *Gibbsian-EM*) even though the β estimate is poorer. The algorithms divide into two groups: on one hand, *MCVEM* and *MC2-EM* which provide lower values of β and consequently images with possibly more isolated points (Figs. 3 and 4); on the other hand, the *Mean Field*, *Simulated Field*, and *Gibbsian-EM* algorithms that provide larger β estimates and smoother images. We observe that the algorithms provide β values larger than the critical value β_c. The later is often considered as a reasonable guess for a fixed β in natural imagery but, as illustrated in Fig. 2, running *MCVEM* and *MC2-EM* with β fixed to β_c (0.88) results in higher error rates. Also it appears clearly, e.g., on the logo image, that *MCVEM* tends to better preserve fine structures, the continuous lines in the original image being less interrupted in various locations (see also the satellite image in Fig. 6). It performs slightly better than *Simulated Field* and *Mean Field*. The triangle image with no such fine structures cannot illustrate this ability of the algorithm. However, we observed the same phenomenon on various other synthetic images with fine structures. On the contrary, when large homogeneous area exists, *MCVEM* and *MC2-EM* segmentations are not smooth enough and isolated points are still visible, producing consequently slightly higher error rates (Fig. 4 and Table V). Note that, in practice, such points are not an issue since they can be easily dealt with afterwards using some simple morphological operator leading to potentially further improved error rates. For example, application of a median filter on the *MCVEM* image reconstruction improves the error rate, 2.73% instead of 2.89% for the logo image (Fig. 3, bottom right), 0.63% instead of 0.81% for the triangle image (Fig. 4, bottom right). Similar conclusions can be drawn from the real image experiments. For the dog lung image (Fig. 5), the *MCVEM* and *MC2-EM* segmentations are not as smooth when considering the light grey region but provide a more accurate segmentation of the white region. For instance, the segmentation of the upper and central parts of the right lung looks better. All spatial algorithms provide, however, smoother

segmentations than *k-means* and *ind-EM*. For the *Simulated Field* algorithm, we report the segmentations corresponding to the implementation of [6] as specified at the end of Section II; however, for this image, 200 iterations are not enough for convergence. We observe that when carrying out more simulations at each iteration or similarly when performing more iterations, the *Simulated Field* algorithm tends to loose the small regions (e.g., the central and background small regions in Fig. 5) so that the segmentations are then very close to the *Gibbsian-EM* ones.

As mentioned earlier, for each algorithm, the displayed segmentation is computed using the current state of the algorithm after a fixed number of iterations. The error rate computed at each iteration, after the burn-in period, stabilizes for *Mean Field* and *MCVEM*. For example, on the logo image of Fig. 3, the error rate is almost always constant. For fixed values of the parameters, the *MCVEM* segmentation procedure does not require simulations any more and is equivalent to the *Mean Field* segmentation procedure. This is not true for the *MC2-EM*, *Gibbsian-EM*, and the *Simulated Field* procedures which remain stochastic since even for fixed values of the parameters, the segmentation step still relies on samples drawn from the conditional field. Nevertheless, for *MC2-EM* and *Gibbsian-EM*, the error rate has a small variation along the path (0.04 for the logo image) while the *Simulated Field* algorithm provides the most unstable procedure since, as already mentioned, its paths do not converge. For the logo image, the error rate variation is 0.11. More complex segmentation rules could be considered to overcome this instability. For example, the different segmentations that can be computed along the iterations can be seen as successive votes, and the final image reconstruction based on the mean value of these votes. For the logo image, this yields for, respectively, the *Simulated Field*, the *Gibbsian-EM* and the *MC2-EM* algorithms a mean error rate of 3.18%, 2.94%, 2.83%, and a lower variation along the path (respectively, 0.0240, 0.0170, and 0.0125).

VI. Discussion and Future Work

In this paper, we proposed a new algorithm to carry out Markov model-based segmentation in practice, combining variational and MCMC ideas. This combination allowed us to prove the first, to our best knowledge, convergence result for this kind of algorithm. This result extends to a whole new class of algorithms. It is based on the idea of seeing the algorithm under study as a perturbed version of a reference algorithm for which convergence results are well established and usually based on a well identified Lyapunov function. For instance, this applies when the model complexity leads to an exact deterministic algorithm which is intractable and must be replaced for practical implementation by an approximate version. The key idea in our contribution, is that although a Lyapunov function does not usually exist for the perturbed algorithm, it is possible to control the distance to this Lyapunov function. Studying its limit set is then made possible through the definition of a set such as $\mathcal{L}$ in Section IV, which defines the algorithm solutions as satisfying an optimality criterion. These observations open the way to a general approach to implement intractable (deterministic) algorithms in practice through adequately designed stochastically perturbed versions. In the hidden Markov random fields context, a natural development of the present work would then be to further study other noisy EM versions with preserved limit sets.

Regarding the *MCVEM* algorithm we focused on, we showed that in addition to guaranteed convergence properties, it provided good segmentation results and compared favorably to other approximated algorithms. Various experiments pointed out that *MCVEM* was close to the *MC2-EM* algorithm based on the *MCEM* algorithm which is known to converge to local maxima of the incomplete log-likelihood. *MCVEM* is then clearly to be favored since it has a much lower computational cost than *MC2-EM*. In particular, the segmentation step in *MCVEM* is simple and does not require the additional computations needed in *MC2-EM*. Also, *MCVEM* tends to provide adequate regularizations through values of β which are not too large and has this way the ability to preserve fine structures. This characteristic can also be responsible for misclassified pixels but they mainly correspond to isolated points. These points can be easily dealt with using some straightforward postprocessing procedure. The performance of *MCVEM* is then very satisfying, all the more so as the results could be further improved by more focus on the use of better sampling techniques. For illustration purpose, we restricted to a simple Gibbs sampler but investigating the use of more sophisticated methods (e.g., [42] and [43]) would be worthwhile. More generally, an alternative approach of the sampling problem would be to consider stochastic approximation techniques such as presented and used in [44] and [45]. We suspect the same kind of convergence results could follow using the same idea of controlling the distance to a reference Lyapunov function.

In this paper, comparison with other existing EM-like procedures showed that the relationship between our algorithm and the former was not obvious. Our study revealed three groups. *MC2-EM* and *MCVEM* distinguish from the *Gibbsian-EM* of [11] and from the *Mean Field* and *Simulated Field* algorithms of [6]. *Simulated Field* does not converge in the same sense and is closer to the *Gibbsian-EM*. It tends to produce smoother segmentations but more unstable trajectories. *Mean Field* has a third specific behavior. Its convergence is not always guaranteed and when observed, the resulting segmentations are very smooth. Further comparisons and investigations would be useful. We believe this first effective step opens the way to a better understanding of the behavior and theoretical properties of a lot of Markov model based algorithms. In particular, analyzing how simulation steps should be incorporated so as to interact advantageously with deterministic approximations seems promising.

Appendix I
Convergence Theorems

A. Model Assumptions

We assume that

A1: $\mathcal{Z}$ is finite, $\mathcal{B} \subseteq \mathbb{R}$, $\Theta \subseteq \mathbb{R}^{n_\theta}$ and $\Psi = \Theta \times \mathcal{B}$.

A2:

i) The function $\psi \mapsto p_{(Y,Z)}(\mathbf{y}, \mathbf{z}; \psi)$ is continuous on Ψ.

ii) For all $q \in \mathcal{I}_r$, the set $\arg\max_{\theta \in \Theta} \sum_{\mathbf{z} \in \mathcal{Z}} p_{Y|Z}(\mathbf{y}|\mathbf{z}; \theta) q(\mathbf{z})$ is not empty.

iii) For any $\mathbf{z} \in \mathcal{Z}$, $\beta \mapsto H(z;\beta)$ is twice-continuously differentiable on $\mathcal{B}$.
The function $\beta \mapsto -\{\sum_{\mathbf{z}\in\mathcal{Z}} H(\mathbf{z};\beta)q(\mathbf{z}) + \ln W(\beta)\}$ is strictly concave and admits a unique maximum in $\mathcal{B}$ for any $q \in \mathcal{I}$.

iv) The function $\beta \mapsto -\{\sum_{\mathbf{z}\in\mathcal{Z}} H(\mathbf{z};\beta)q(\mathbf{z}) + \ln \tilde{W}^{J,b}(\beta)\}$ is strictly concave and admits a unique maximum in $\mathcal{B}$, for any $q \in \mathcal{I}_r$, any integer J and any $b \in \mathcal{B}$.

Define the function L on $\mathcal{I} \times \Psi$ by $L(q,\psi) = \exp(F(q,\psi))$where F is given by (5).

A3: For any $M > 0$, the level set $\{(q,\psi) \in \mathcal{I}_r \times \Psi, L(q,\psi) \geq M\}$ is bounded.

A4: Assume either that *i)* the set $L(\mathcal{L})$ is compact, or *ii)* for all compact $\mathcal{K} \subset \mathcal{I}_r \times \Psi$, $L(\mathcal{K} \cap \mathcal{L})$ is finite, where $\mathcal{L}$ is defined by (11).

Under **A1** and **A2i)**, L is a continuous function and the level set is a closed subset of $\mathcal{I}_r \times \Psi$. Hence, it is compact in $\mathcal{I}_r \times \Psi$. Furthermore, $\mathcal{L}$ is a closed subset of $\mathcal{I}_r \times \Psi$. Hence, **A4** is satisfied whenever $L(\mathcal{L})$ is bounded.

B. Monte Carlo Approximations

We formulate sufficient conditions that imply a local uniform control of the difference between the gradient $\nabla \ln W$ and its Monte Carlo approximation. Let $\mathbb{E}_{\lambda,\beta}$ be the expectation on the canonical space associated to the Markov chain with initial distribution λ and stationary distribution $p_Z(\cdot;\beta)$. Let $\mathrm{Cl}(\mathcal{C}_\alpha)$ be the closure of the α-neighborhood of some (bounded) set $\mathcal{C}$.

A5: There exist $r \geq 2$ and a probability distribution λ on $\mathcal{Z}$ such that for any compact subset $\mathcal{C} \subset \mathcal{B}$ and any $\alpha > 0$

$$\sup_{\beta\in\mathrm{Cl}(\mathcal{C}_\alpha),b\in\mathcal{C}} \sup_{J\geq 1} J^{r/2}\mathbb{E}_{\lambda,b}\left[\left|\nabla_\beta\left\{\ln \tilde{W}^{J,b}(\beta) - \ln W(\beta)\right\}\right|^r\right]$$

is finite.

A5 is satisfied whenever

$$\sup_{\beta\in\mathrm{Cl}(\mathcal{C}_\alpha),b\in\mathcal{C}} \sup_{J\geq 1} J^{r/2}\mathbb{E}_{\lambda,b}\left[\left|\tilde{W}^{J,b}(\beta) - W(\beta)\right\}\right|^r\right]$$

is finite and

$$\sup_{\beta\in\mathrm{Cl}(\mathcal{C}_\alpha),b\in\mathcal{C}} \sup_{J\geq 1} J^{r/2}\mathbb{E}_{\lambda,b}\left[\left|\nabla_\beta\left\{\tilde{W}^{J,b}(\beta) - W(\beta)\right\}\right|^r\right]$$

is finite. Observe that both of these integrals are on the form

$$\mathbb{E}_{\lambda,b}\left[\left|\sum_{j=1}^{J}\left\{\mathcal{H}(Z^j;\beta,b) - \sum_{\mathbf{z}}\mathcal{H}(\mathbf{z};\beta,b)p_Z(\mathbf{z};b)\right\}\right|^r\right]$$

where $p_Z(\mathbf{z};b)$ is the invariant probability distribution of the Markov chain $\{Z^j\}_j$ with initial distribution λ. Sufficient conditions implying this uniform control of the L^r-norm difference for a Markov chain can be found in [7] (see Section V for an example). Finally, we assume that the number of simulations $\{J_t\}_t$ increases at a rate such that the larger r, the weaker the rate.

A6: $\{J_t\}_t$ is a positive integer-valued sequence such that $\sum_{t\geq 0} J_t^{-r/2} < \infty$ where r is given by **A5**.

C. Convergence Theorems

Consider the generalized VEM algorithm that replaces in the β-update, the global optimization by a local one on $\mathcal{B}^t$. The proof of Theorem 1 is along the same lines as the proof of [4, Theorem 2(i)] (see also [36]) and is, thus, omitted.

Theorem 1: Assume **A1**, **A2i)–A2iii)** and **A3**. Fix a positive sequence $\{\gamma^t\}_t$ such that $\inf_t \gamma^t > 0$ and let $\{(q^t,\psi^t)\}_t$ be the generalized VEM path started at $(q^0,\psi^0) \in \mathcal{I}_r \times \Psi$.

The sequence $\{L(q^t,\psi^t)\}_t$ converges monotonically to $L^* = L(q^*,\psi^*)$ for some $(q^*,\psi^*) \in \mathcal{L}$. Furthermore, the sequence $\{(q^t,\psi^t)\}_t$ converges to the set $\{(q,\psi) \in \mathcal{L}, L(q,\psi) = L^*\}$.

The convergence of the random trajectories of MCVEM is established almost surely with respect to $\bar{\mathbb{P}}$, the probability on the canonical space associated to the trajectories started at (q^0, ψ^0), given the initial distribution λ of the Markov chain, the sequence of compact sets $\{\mathcal{C}^t\}_t$ satisfying (9) and the sequence $\{\gamma^t\}_t$.

Theorem 2: Assume **A1**–**A6**. Let $\{\mathcal{C}^t\}_t$ be a sequence of compact sets satisfying (9), $(q^0,\psi^0) \in \mathcal{I}_r \times \mathcal{C}^0$ and λ be given in **A5**. Fix a positive sequence $\{\gamma^t\}_t$ such that $\inf_t \gamma^t > 0$. Consider the MCVEM random sequence $\{(q^t,\psi^t)\}_t$. Then, $\lim_t \tau^t < \infty$ w.p.1 and $\limsup_t |\psi^t| < \infty$ w.p.1. and $\{L(q^t,\psi^t)\}_t$ converges w.p.1 to a connected component of $L(\mathcal{L})$. If in addition $L(\mathcal{L} \cap \mathrm{Cl}(\{(q^t,\psi^t)\}_t))$ has an empty interior, then $\{L(q^t,\psi^t)\}_t$ converges w.p.1 to L^* and $\{(q^t,\psi^t)\}_t$ converges to the set $\mathcal{L}_{L^*} = \{(q,\psi) \in \mathcal{L}, L(q,\psi) = L^*\}$.

Observe that if **A4*ii)*** is satisfied, $L(\mathcal{L} \cap \mathrm{Cl}(\{(q^t,\psi^t)\}_t))$ is finite, thus having an empty interior. The proof of Theorem 2 is very close to the proof of [7, Theorem 3]. The first step consists in an extension of some deterministic results ([7, Propositions 9, 10, 11]) in order to take into account that in the present case, any MCVEM iteration corresponds to an inhomogeneous point-to-set map (in [7], only point-to-point maps are addressed). These deterministic results provide sufficient conditions for convergence of some iterated perturbed map that approximates, in the sense given by (12), an iterative map having a Lyapunov function. Convergence of MCVEM then results from an application of these propositions. The most technical step is to prove that (12) holds $\bar{\mathbb{P}}$—a.s. It is sufficient to prove that for all $\epsilon > 0$, the random series with general term $\mathbb{1}_{\{|L(q^{t+1},\psi^{t+1}) - L(q^{t+1},\bar{\psi}^{t+1})|\mathbb{1}_{(q^t,\psi^t)\in\mathcal{K}} \geq \epsilon\}}$ is finite a.s. Following the same lines as in the proof of [7, Theorem 3], this series is finite whenever, for some $\alpha > 0$ depending upon ϵ, $\sum_{t\geq 0} \bar{\mathbb{P}}(|\bar{\beta}^{t+1} - \beta^{t+1}| \geq \alpha|\mathcal{F}_t,)\mathbb{1}_{(q^t,\psi^t)\in\mathcal{K}}$ is finite a.s. which is true under **A5**. Here, $\mathcal{F}_t$ is the σ-field that contains the random variables $\{Z^{j,s}, j \leq J_s, s \leq t-1\}$, $\bar{\beta}^{t+1}$ is the maximum of $\beta \mapsto -\{\sum_{\mathbf{z}\in\mathcal{Z}} H(\mathbf{z};\beta)q^{t+1}(\mathbf{z}) + \ln W(\beta)\}$ over $\mathcal{B}^t$, β^{t+1} is the maximum of $\beta \mapsto -\{\sum_{\mathbf{z}\in\mathcal{Z}} H(\mathbf{z};\beta)q^{t+1}(\mathbf{z}) + \ln \tilde{W}^{J_t,\beta^t}(\beta)\}$ over $\mathcal{B}^t$, and q^{t+1} is given by Step i) of MCVEM. The interested reader will refer to [35] for a detailed proof.

APPENDIX II
APPLICATION TO IMAGE SEGMENTATION

We show that the model described in Section V satisfies the conditions **A1**–**A6**. We assumed that at each pixel, the observations are univariate; this is not at all restrictive and the multidimensional case could be considered in the same way.

Conditions **A1** and **A2** are easily checked; for **A2**, we use the strict concavity of $\beta \mapsto \ln W(\beta)$. Details are omitted. Regarding assumption **A3**, since L is a continuous positive function and $\mathcal{I}_r$ is bounded it is enough to show that for $q \in \mathcal{I}_r$, L tends to 0 on the boundaries of Ψ. Let $\ln L$ be divided in three parts, $\ln L(q,\psi) = a(q,\theta) + b(q,\beta) + c(q)$, where (up to an additive constant independent of the parameters)

$$a(q,\theta) = -\frac{1}{2}\sum_{i\in S}\sum_{k=1}^{K}\left[\ln(\sigma_k) + \sigma_k^{-2}(y_i-\mu_k)^2\right] q_i(e_k),$$
$$b(q,\beta) = \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \ln\left(p_Z(\mathbf{z};\beta)\right).$$

For $q \in \mathcal{I}_r$, for all $l = 1,\ldots,K$ there exist $i \in S$ such that $q_i(e_l) > 0$. Then whenever there exists k such that σ_k tends to 0 or μ_k tends to $\pm\infty$, part $\sigma_k^{-2}(y_i-\mu_k)^2$ is the most significant term in expression $a(q,\theta)$. If σ_k tends to $+\infty$, then the most significant term in $a(q,\theta)$ is $\ln(\sigma_k)$. In all cases $a(q,\theta)$ tends to 0. When β tends to $+\infty$ (respectively, $-\infty$), then $p_Z(\mathbf{z};\beta)$ tends to 0 except in $\mathbf{z} \in \arg\max_z h(\mathbf{z})$ (respectively, $\mathbf{z} \in \arg\min_z h(\mathbf{z})$) so that clearly $b(q,\beta)$ tends to 0. It follows that **A3** is satisfied.

For **A4**, we show that $\mathcal{L}$ is compact which implies that $L(\mathcal{L})$ is compact since L is continuous. Under the stated assumptions, $\mathcal{L}$ is closed and it remains to prove that $\mathcal{L}$ is bounded. Let us first observe that for $(q^*,\psi^*) \in \mathcal{L}$, q^* is included in a compact set and ψ^* satisfies $\nabla_\psi F(q^*,\psi^*) = 0$, which leads to closed-form expressions

$$\forall k, \quad \mu_k^* = \frac{\sum_{i\in S} y_i q_i^*(e_k)}{\sum_{i\in S} q_i^*(e_k)}$$
$$\sigma_k^{2*} = \frac{\sum_{i\in S}(y_i-\mu_k^*)(y_i-\mu_k^*)^t q_i^*(e_k)}{\sum_{i\in S} q_i^*(e_k)}$$

hence, q^* and θ^* are linked through a continuous and bounded function on $\mathcal{I}_r$ and θ^* is bounded. By applying the implicit function theorem we prove that the same holds for β^* which shows that $\mathcal{L}$ is bounded.

For **A5**, we can actually show that a more general condition holds: applying the results by [7], we can deduce that the conditions in **A5** hold for all $r \geq 2$ and any initial distribution λ. Referring to [7, Proposition 1], it is enough to show that for the Markov chain used in the approximation of $W(\beta)$, the state space is *small* (see, e.g., [46]). The Gibbs sampler is a Markov chain with kernel $P = P_1 P_2 \ldots P_N$ where P_n replaces the n^{th} pixel with a draw from the conditional $p_Z(z_n|\mathbf{z}_{S\setminus\{n\}})$ leaving $\mathbf{z}_{S\setminus\{n\}}$ unchanged. Since $\mathcal{Z} = V^N$ is a product space with V finite, the small set property follows.

ACKNOWLEDGMENT

The authors would like to thank J. Blanchet for help with some experiments, and M. Sigelle and O. Cappé for fruitful discussions.

REFERENCES

[1] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA: MIT Press, 1998, pp. 105–162.
[2] M. Wainwright and M. Jordan, "A variational principle for graphical models," in *New Directions in Statistical Signal Processing*, S. Haykin, T. Principe, T. Sejnowski, and J. McWhirter, Eds. Cambridge, U.K.: MIT Press, 2005, ch. 11.
[3] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 1999.
[4] W. Byrne and A. Gunawardana, "Convergence theorems of generalized alternating minimization procedures," *J. Mach. Learn. Res.*, vol. 6, pp. 2049–2073, 2005.
[5] J. Zhang, "The mean fields theory in EM procedures for Markov random fields," *IEEE Trans. Signal Process.*, vol. 40, no. 10, pp. 2570–2583, Oct. 1992.
[6] G. Celeux, F. Forbes, and N. Peyrard, "EM procedures using mean-field approximations for Markov model-based image segmentation," *Pattern Recognit.*, vol. 36, pp. 131–144, 2003.
[7] G. Fort and E. Moulines, "Convergence of the Monte-Carlo EM for curved exponential families," *Ann. Statist.*, vol. 31, no. 4, pp. 1220–1259, 2003.
[8] G. Wei and M. Tamier, "A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm," *J. Amer. Statist. Assoc.*, vol. 85, pp. 699–704, 1990.
[9] N. de Freitas, P. Hojen-Sorensen, M. Jordan, and S. Russel, "Variational MCMC," Tech. Rep., Univ. California, Berkeley, 2001.
[10] W. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds., *Markov Chain Monte Carlo in Practice*. Suffolk, U.K.: Chapman & Hall, 1996.
[11] B. Chalmond, "An iterative Gibbsian technique for reconstruction of m-array images," *Pattern Recognit.*, vol. 22, no. 6, pp. 747–761, 1989.
[12] W. Qian and D. Titterington, "Estimation of parameters in hidden Markov models," *Philos. Trans. Roy. Soc. Lond. A*, vol. 337, pp. 407–428, 1991.
[13] ——, "Discussion of a paper by Geyer and Thompson," *J. Roy. Statist. Soc. B*, vol. 54, pp. 657–699, 1992.
[14] J. Archer and D. Titferington, "Parameter estimation for hidden markov chains," *J. Statist. Planning Inference*, vol. 108, pp. 365–390, 2002.
[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.
[16] C. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.
[17] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures," *Statist. Dec.*, no. 1, pp. 205–237, 1984.
[18] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse and other variants," in *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA: MIT Press, 1998, pp. 355–368.
[19] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1996.
[20] R. Boyles, "On the convergence of EM algorithms," *J. Roy. Statist. Soc B*, vol. 45, no. 1, pp. 47–50, 1983.
[21] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," Tech. Rep. 649, Dept. Statist., Univ. California, Berkeley, 2003.
[22] T. Tanaka, "Information geometry of mean-field approximation," in *Advanced Mean Field Methods*, M. Opper and D. Saad, Eds. Cambridge, MA: MIT Press, 2001, ch. 17.
[23] C.-H. Wu and P. C. Doerschuk, "Cluster expansions for the deterministic computation of Bayesian estimators based on Markov random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 3, pp. 275–293, Mar. 1995.
[24] C. Ambroise and G. Govaert, "Convergence proof of an EM-type algorithm for spatial clustering," *Pattern Recognit. Lett.*, vol. 19, pp. 919–927, 1998.
[25] J. Zhang, "The convergence of mean field procedures for MRF's," *IEEE Trans. Image Process.*, vol. 5, no. 12, pp. 1662–1665, Dec. 1996.
[26] J. Fessler, "Comments on "The convergence of mean field procedures for MRF's"," *IEEE Trans. Image Process.*, vol. 7, no. 6, p. 917, Jun. 1998.

[27] C. Geyer and E. Thompson, "Constrained Monte-Carlo maximum likelihood for dependent data (with discussion)," *J. Roy. Statist. Soc. B*, vol. 54, pp. 657–699, 1992.
[28] X. Desconibes, R. Morris, J. Zerubia, and M. Berthod, "Estimation of Markov random field prior parameters using Markov Chain Monte Carlo maximum likelihood," *IEEE Trans. Image Process.*, vol. 8, no. 7, pp. 954–963, Jul. 1999.
[29] J. Geweke, "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, vol. 24, pp. 1317–1399, 1989.
[30] L. Younes, "Monte-Carlo maximization of likelihood: A convergence study," Tech. Rep., CMLA, ENS Cachan, France, 2000 [Online]. Available: http://www.cmla.ens-cachan.fr/UtiIisateurs/younes
[31] S. Gemao and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.
[32] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chetn. Phys.*, vol. 21, pp. 1087–1091, 1953.
[33] R. Swendsen and J. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Phys. Rev. Lett.*, vol. 58, pp. 86–88, 1987.
[34] A. Gelman and X.-L. Meng, "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling," *Statist. Sci.*, vol. 13, pp. 163–185, 1998.
[35] F. Forbes and G. Fort, "A convergence theorem for variational EM-like algorithms: Application to image segmentation," Tech. Rep. RR-572I, INRIA, Rhône-Alpes, France, 2005 [Online]. Available: http://www.inria.fr/rrrt/rr-5721.html
[36] W. Zangwill, *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1969.
[37] J. Maroquin, S. Milter, and T. Pogio, "Probabilistic solution of ill-posed problems in computer vision," *J. Ann. Statist. Assoc.*, vol. 82, pp. 76–89, 1987.
[38] S. Nielsen, "The stochastic EM: estimation and asymptotic results," *Bernoulli*, vol. 6, no. 3, pp. 457–489, 2000.
[39] G. Celeux and J. Diebolt, "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Comput. Statist. Quart.*, vol. 2, no. 1, pp. 73–82, 1985.
[40] B. Potyak and A. Juditski, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, 1992.
[41] D. Stanford, "Fast automatic unsupervised image segmentation and curve detection in spatial point processes," Ph.D. dissertation, Dept. Statist., Univ. Washington, Seattle, 1999.
[42] C. Sminchisescu, M. Welling, and G. Hinton, "A mode-hopping MCMC sampler," Tech. Rep. CSRG-478, Univ. Toronto, Toronto, ON, Canada, 2003.
[43] A. Frigessi, C.-R. Hwang, and L. Younes, "Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields," *Ann. Appl. Probab.*, vol. 2, no. 3, pp. 610–628, 1992.
[44] L. Younes, "Parametric inference for imperfectly observed Gibbsian fields," *Probab. Theory Rel. Fields*, vol. 82, pp. 625–645, 1989.
[45] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the EM algorithm," *Ann. Statist.*, vol. 27, no. 1, pp. 94–128, 1999.
[46] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*. London, U.K.: Springer-Verlag, 1993.

Florence Forbes was born in Monaco in 1970. She received the M.Sc. degree in computer science and applied mathematics from the Ecole National Supérieure Informatique et Mathématiques Appliquées de Grenoble, Grenoble, France, in 1993, and the Ph.D. degree in applied probabilities in 1996 from the University Joseph Fourier, Grenoble.

She is a Research Scientist at the Institut National de Recherche en Informatique et Automatique (INRIA), Montbonnot, France. She joined the IS2 research team at INRIA Rhône-Alpes, France, in 1998, where she has been Head of the MISTIS team since 2003. Her research activities include Bayesian image analysis, Markov models, and hidden structure models.

Gersende Fort was born in Rouen, France, in 1974. She received the M.Sc. degree in telecommunications engineering from the Ecole Nationale Supérieure des Téelécommunications, Paris, France, in 1997, and the Ph.D. degree in applied mathematics from University Pierre et Marie Curie (Paris VI), Paris, in 2001.

She is currently with the Centre National de la Recherche Scientifique (CNRS) at ENST/ LTCI, Paris. Her research interests are in simulation methods by Markov chain Monte Carlo and in computational statistics. Her current professional activities include participating in the "Adaptive Monte Carlo Methods" program at the National Research Agency, and she is a member of the research team MISTIS of the Institut National de Recherche en Informatique et Automatique, Montbonnot, France.