EM procedures using mean field-like approximations for Markov model-based image segmentation

Gilles Celeux Florence Forbes *
Nathalie Peyrard
Inria Rhône-Alpes †

December 11, 2001

*Correspondence should be sent to Florence Forbes: Fax: (00 33) 4 76 61 54 77, Florence.Forbes@inrialpes.fr

†Gilles Celeux and Florence Forbes are Researchers, Nathalie Peyrard is PhD student, Projet IS2, Inria Rhône-Alpes, Zirst, 655 av. de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France. Emails: Gilles.Celeux@inrialpes.fr, Florence.Forbes@inrialpes.fr, Nathalie.peyrard@inrialpes.fr.

Abstract: Image segmentation using Markov random fields involves parameter estimation in hidden Markov models for which the EM algorithm is widely used. In practice, difficulties arise due to the dependence structure in the models and approximations are required. Using ideas from the Mean Field approximation principle, we propose a class of EM-like algorithms in which the computation reduces to dealing with systems of independent variables. Within this class, the *simulated field algorithm* is a new stochastic algorithm which appears to be the most promising for its good performance and speed, on synthetic and real image experiments.

Key Words: Image segmentation, Hidden Markov random fields, EM algorithm, ICM algorithm, Pseudo-likelihood, Mean field approximation, Simulated field.

1 Introduction

Hidden Markov random field (HMRF) models revealed themselves as a powerful tool for image segmentation [1], [2]. They are very useful in accounting for spatial dependences between the different pixels of an image but these spatial dependences are also responsible for a typically large amount of computation. In practice, Markov model-based segmentation requires estimation of the model parameters. A common approach consists of alternatively restoring the unknown segmentation based on a Maximum A Posteriori (MAP) rule and then estimating the model parameters using the observations and the restored data. This is the case, for instance, in the popular ICM algorithm of [2] which makes use of the pseudo-likelihood approximation [3]. This combination usually provides reasonable segmentations but is known to lead to biased parameter estimates, essentially due to the restoration step. In this work, we consider the Expectation-Maximization (EM) algorithm which uses conditional probabilities rather than restorations of the hidden data. This procedure is widely used in the context of incomplete data and in particular to estimate independent mixture models due to its simplicity. As regards HMRF, the model complexity is much greater and makes the algorithm intractable. Solutions have been proposed which associate the pseudo-likelihood approximation [3] and Monte Carlo simulations [4]. But the corresponding algorithms are time consuming. In this paper, we propose a generalization of the mean field principle of statistical physics [5] to make the EM algorithm tractable. More specifically, we consider approximations of Markov models, with complex dependences, by systems of independent variables. These approximations are obtained by fixing the neighborhood of each pixel to arbitrary constants. They lead to valid probability models, with factorization properties, much simpler to deal with. We then use these approximations to carry out the EM algorithm and derive a class of algorithms. They can be interpreted as the EM algorithm for independent mixture models with the difference that the mixture model adaptatively changes at each iteration depending on the current choice for the pixels' neighbors labels. It follows algorithms that have the advantage to take the spatial information into account while keeping each iteration as simple as in the independent mixture case. The class of algorithms that we exhibit includes new and already known procedures (*e.g.* ICM and the *mean field* algorithm of [6]), providing a unified and insightful view of various HMRF parameter estimation procedures. Within this class a new stochastic algorithm, the *simulated field algorithm*, deserves to be distinguished. At each iteration, the true model is approximated by setting the neighborhood of each pixel to simulated values from a distribution approximating the relevant untractable Markov distribution. In most cases, it performs better than the other algorithms in its class. In particular, it can shows significant improvements when compared to Zhang's mean field algorithm ([6], [7]), the ICM algorithm ([2]) and the EM algorithm for independent mixture models. This confirms its ability to preserve the good features of EM while making significant use of the spatial information in a reasonable amount of time.

The paper is organized as follows. Section 2 specifies the context of hidden Markov models for image segmentation. The mean field approximation principle is presented in Section 3. The proposed class of algorithms is described and analysed in Section 4. In Section 5, we consider other algorithms and present them in our unified framework. The performance of several algorithms derived from our approach are illustrated on synthetic and real images in Section 6 and a discussion section ends the paper.

2 Markov model-based image segmentation

Problems involving incomplete data, where part of the data is missing or unobservable, are common in image analysis. The aim may be to recover an original image which is hidden and has to be estimated from a noisy or blurred version. More generally, the observed and hidden data are not necessarily of the same nature. The observations may represent measurements, *e.g.* multidimensional variables recorded for each pixel of an image while the hidden data could consist of an unknown class assignment to be estimated from the observations for each pixel. This case is usually referred to as image segmentation. In this paper, we focus on Markov model-based image segmentation. In Section 2.1, we recall basic definitions concerning the Markov models used for the unobserved data. In Section 2.2, we specify the complete parametric models for the observed and unobserved data. In Section 2.3, we describe the parameter estimation algorithm we considered for these models.

2.1 Markov random fields and pseudo-likelihood approximation

Let S be a finite set of sites with a neighborhood system defined on it. Let $|S|$ denote the number of sites. A typical example in image analysis is the two dimensional lattice with a second order neighborhood system. For each site, the neighbors are the eight 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 $\{0,1\}^{K\times K}$. 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 satisfies the following properties,

$$\forall \mathbf{z}, \quad P_G(z_i \mid \mathbf{z}_{S\backslash\{i\}}) \;=\; P_G(z_i \mid z_j,\ j \in N(i)) \tag{1}$$

$$\forall \mathbf{z}, \quad P_G(\mathbf{z}) \;>\; 0, \tag{2}$$

where $\mathbf{z}_{S\backslash\{i\}}$ denotes a realization of the field restricted to $S\backslash\{i\} = \{j \in S, j \neq i\}$ and $N(i)$ denotes the set of neighbors of i. More generally, if A is a subset of S, we will write $\mathbf{z}_A$ for $\{z_i,\ i \in A\}$. In words, property (1) means that the interactions between site i and the other sites actually reduce to interactions with its neighbors. Property (2) is important for the Hammersley-Clifford theorem to hold. This theorem states that the joint probability distribution of a Markov field is a Gibbs distribution (for which we use the notation P_G) given by

$$P_G(\mathbf{z}) \quad = \quad W^{-1} \exp(-H(\mathbf{z})), \tag{3}$$

where H is the energy function

$$H(\mathbf{z}) \quad = \quad \sum_c V_c(\mathbf{z}_c). \tag{4}$$

The V_c's are the clique potentials and 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. We will write $\sum_{\mathbf{z}}$ (resp. $\sum_{\mathbf{z}_A}$) a sum over all possible values of $\mathbf{z}$ (resp. $\mathbf{z}_A$). 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 a problem when using these models in situations where an expression of the joint distribution $P_G(\mathbf{z})$ is required. An approximation of the likelihood (3) is the pseudo-likelihood introduced by [3] and defined as

$$\mathcal{PL}(\mathbf{z}) \quad = \quad \prod_{i \in S} P_G(z_i \mid \mathbf{z}_{N(i)}). \tag{5}$$

Each term in the product is easy to compute,

$$P_G(z_i \mid \mathbf{z}_{N(i)}) \quad = \quad \frac{\exp(-\sum_{c \ni i} V_c(\mathbf{z}_c))}{\sum_{z_i} \exp(-\sum_{c \ni i} V_c(\mathbf{z}_c))}.$$

Expression (5) is a genuine probability distribution only when the variables are independent but it can be used to obtain estimates of a Markov random field parameters. In Sections 3 and 4, we will propose other approximations based on systems of independent variables. Their factorization properties simplify computations as (5) and they correspond to valid probability models.

2.2 Hidden Markov models

Image segmentation involves observed data and unobserved data to be recovered. In this paper, the unobserved data is modeled as a discrete Markov random field, $\mathbf{Z}$, as defined in (3) with energy function H depending on a parameter β. In hidden Markov models, the observations $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$, according to a density f which is assumed to be of the following type (θ is a parameter and the f_i's are given),

$$\begin{aligned} f(\mathbf{y} \mid \mathbf{z},\, \theta) &= \prod_{i \in S} f_i(y_i \mid z_i,\, \theta) \qquad (6) \\ &= \exp\{\sum_{i \in S} \log f_i(y_i \mid z_i,\, \theta)\}, \end{aligned}$$

assuming that all the $f_i(y_i \mid z_i,\, \theta)$ are positive. This makes the model similar to an independent mixture model (*cf.* [8]). An independent mixture model could be seen as a hidden Markov model where the hidden field $\mathbf{Z}$ is one of independent identically distributed variables. In the general case, the complete likelihood is given by

$$\begin{aligned} P_G(\mathbf{y}, \mathbf{z} \mid \theta, \beta) &= f(\mathbf{y} \mid \mathbf{z}, \theta)\; P_G(\mathbf{z} \mid \beta) \\ &= W(\beta)^{-1} \prod_{i \in S} f_i(y_i \mid z_i,\, \theta)\; \prod_c \exp\{-V_c(\mathbf{z}_c \mid \beta)\} \qquad (7) \\ &= W(\beta)^{-1} \exp\{-H(\mathbf{z} \mid \beta) + \sum_{i \in S} \log f_i(y_i \mid z_i,\, \theta)\}. \end{aligned}$$

Thus the conditional field $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ is a Markov field as $\mathbf{Z}$ is. Its energy function is

$$H(\mathbf{z} \mid \mathbf{y}, \theta, \beta) \;=\; H(\mathbf{z} \mid \beta) - \sum_{i \in S} \log f_i(y_i \mid z_i,\, \theta).$$

In the following developments, we will refer to Markov fields $\mathbf{Z}$ and $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ as the marginal and 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, an estimation of Ψ can be obtained via the EM algorithm that we describe below.

2.3 Parameter estimation using the EM algorithm

Assuming Ψ unknown, our aim is to get the maximun likelihood estimate of this parameter knowing the observations $\mathbf{y}$. The log-likelihood of the model is

$$L(\Psi) \;=\; \log P_G(\mathbf{y} \mid \Psi) \;=\; \log \sum_{\mathbf{z}} P_G(\mathbf{y}, \mathbf{z} \mid \Psi).$$

The EM algorithm [9] is an iterative algorithm aiming at maximizing this log-likelihood by maximizing at iteration q,

$$Q(\Psi \mid \Psi^{(q)}) \;=\; \mathbb{E}_{\Psi^{(q)}}[\log P_G(\mathbf{y}, \mathbf{Z} \mid \Psi) \mid \mathbf{Y} = \mathbf{y}] \,,$$

the expectation of the complete log-likelihood knowing the observation $\mathbf{y}$ and current estimate $\Psi^{(q)}$. The EM algorithm can therefore be described as follows,

(1) start from an initial guess $\Psi^{(0)}$ for Ψ,

(2) update the current estimate $\Psi^{(q)}$ to

$$\Psi^{(q+1)} \;=\; \arg\max_{\Psi} Q(\Psi \mid \Psi^{(q)}).$$

The updating part (2) can be divided in two steps. The computation of $Q(\Psi \mid \Psi^{(q)})$ corresponds to the E (expectation) step and the maximization with respect to Ψ to the M (maximization) step.

A well known property of the algorithm is that $L(\Psi^{(q)})$ increases with q. Using (7), Q can be further written as follows

$$\begin{aligned} Q(\Psi \mid \Psi^{(q)}) \;&=\; \sum_{i \in S} \sum_{z_i} P_G(z_i \mid \mathbf{y}, \Psi^{(q)}) \, \log f_i(y_i \mid z_i, \theta) \\ &\quad - \log W(\beta) - \sum_{c} \sum_{\mathbf{z}_c} V_c(\mathbf{z}_c \mid \beta) \, P_G(\mathbf{z}_c \mid \mathbf{y}, \Psi^{(q)}). \end{aligned} \tag{8}$$

The first term does not depend on β while the two last ones do not involve θ. Therefore we will write

$$\begin{aligned} Q(\theta \mid \Psi^{(q)}) \;&=\; \mathbb{E}_{\Psi^{(q)}}[\log P_G(\mathbf{y} \mid \mathbf{Z}, \theta) \mid \mathbf{Y} = \mathbf{y}] \\ &=\; \sum_{i \in S} \sum_{z_i} P_G(z_i \mid \mathbf{y}, \Psi^{(q)}) \, \log f_i(y_i \mid z_i, \theta) \end{aligned} \tag{9}$$

$$\begin{aligned} Q(\beta \mid \Psi^{(q)}) \;=\; \mathbb{E}_{\Psi^{(q)}}[\log P_G(\mathbf{Z} \mid \beta) \mid \mathbf{Y} = \mathbf{y}] \;&=\; -\log W(\beta) \\ &\quad - \sum_{c} \sum_{\mathbf{z}_c} V_c(\mathbf{z}_c \mid \beta) \, P_G(\mathbf{z}_c \mid \mathbf{y}, \Psi^{(q)}). \end{aligned} \tag{10}$$

There are two difficulties in evaluating Q in this case. Both the partition function $W(\beta)$ and the conditional probabilities, $P_G(z_i \mid \mathbf{y}, \Psi^{(q)})$ and $P_G(\mathbf{z}_c \mid \mathbf{y}, \Psi^{(q)})$, cannot be computed exactly. The partition function problem can be solved by using the pseudo-likelihood (5) in place of $P_G(\mathbf{z} \mid \beta)$. This changes the expression in (7) into

$$P_G(\mathbf{y}, \mathbf{z} \mid \Psi) \;\approx\; \prod_{i \in S} \{f_i(y_i \mid z_i, \theta)\; P_G(z_i \mid \mathbf{z}_{N(i)}, \beta)\},$$

so that

$$Q(\Psi \mid \Psi^{(q)}) \;\;\approx\;\; \sum_{i \in S} I\!E_{\Psi^{(q)}}[\log f_i(y_i \mid Z_i, \theta) \mid \mathbf{Y} = \mathbf{y}] + \sum_{i \in S} I\!E_{\Psi^{(q)}}[\log P_G(Z_i \mid \mathbf{Z}_{N(i)}, \beta) \mid \mathbf{Y} = \mathbf{y}].$$

The first sum is expression (9). In the second sum, computing further the expectation leads to

$$Q(\beta \mid \Psi^{(q)}) \;\;\approx\;\; \sum_{i \in S} \sum_{\mathbf{z}_{\overline{N(i)}}} P_G(\mathbf{z}_{\overline{N(i)}} \mid \mathbf{y}, \Psi^{(q)})\; \log P_G(z_i \mid \mathbf{z}_{N(i)}, \beta), \tag{11}$$

where $\overline{N(i)} = N(i) \cup \{i\}$. Then, the conditional probabilities $P_G(z_i \mid \mathbf{y}, \Psi^{(q)})$ in (9) and $P_G(\mathbf{z}_{\overline{N(i)}} \mid \mathbf{y}, \Psi^{(q)})$ in (11) can be approximated using Markov Chain Monte Carlo (MCMC) simulations (see [4]). This requires a large amount of computation. An alternative to these approximations is to use the mean field approximation for both the marginal field $P_G(\mathbf{z} \mid \beta)$ and the conditional field $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$, (see [6] and [10]). The mean field approximation principle is presented in Section 3. In Section 4, we show how to use mean field-like approximations to make the EM algorithm more tractable.

3 Mean field approximation principle

The mean field approximation is originally a method of approximation for the computation of the mean of a Markov random field. It comes from statistical mechanics (*e.g.* [5]) where it has been used as an analysis tool to study phase transition phenomena. More recently, it has been used in computer vision applications (*e.g.* [11], [12], [13]), graphical models (*e.g.* [14] and references therein) and other areas (*e.g.* [15]). It can also be used to provide an approximation of the distribution of a Markov random field. The idea when considering a particular site i is to neglect the fluctuations of the sites interacting with i. The resulting system behaves as one composed of independent variables for which computation gets tractable. More specifically, for all j different from i, the Z_j's are fixed to their mean value $\mathbb{E}_G(Z_j)$, denoted by m_j for all $j \in S\backslash\{i\}$. Let $\mathbf{m}$ denote $\{m_i, i \in S\}$. Replacing the z_j's by their mean value m_j, in expression (4), leads to the definition of a new energy function for site i,

$$H_i^{mf}(z_i) \;=\; H(\mathbf{z})|_{z_j=m_j, j\neq i} \;=\; H(z_i \mathbf{m}_{S\backslash\{i\}}). \tag{12}$$

Notation $z_i\mathbf{m}_{S\backslash\{i\}}$ denotes the configuration equal to z_i at site i and to $\mathbf{m}_{S\backslash\{i\}}$ on $S\backslash\{i\}$. Note that in order to apply the mean field theory, it must be that H which is originally defined on $V^{|S|} \subset \{0,1\}^{K\times|S|}$, can be extended to $[0,1]^{K\times|S|}$. It is also convenient to allow the Z_j, for j in $S\backslash\{i\}$, to take values in $[0,1]^K$ rather than V. It naturally follows the definition of a probability measure, denoted by P_i^{mf}, that concentrates on the manifold $\{Z_j = m_j\}$ for all sites j different from i,

$$P_i^{mf}(\mathbf{z}) \;=\; {W_i^{mf}}^{-1} \exp(-H_i^{mf}(z_i)) \prod_{j\neq i} \delta_{\{Z_j=m_j\}}(z_j),$$

where $W_i^{mf} \;=\; \sum\limits_{z_i} \exp(-H_i^{mf}(z_i))$ and $\delta_A(.)$ is the indicator function of set A. In expression (4), the terms that involve z_i can be isolated from the others. It leads to the decomposition of the mean field energy at pixel i, (12), into the mean field local energy at pixel i, denoted by $H_i^{mfloc}(z_i)$, and a term, $R_i^{mfloc}(\mathbf{m}_{S\backslash\{i\}})$, that does not depend on z_i, namely

$$H_i^{mf}(z_i) \;=\; H_i^{mfloc}(z_i) + R_i^{mfloc}(\mathbf{m}_{S\backslash\{i\}}). \tag{13}$$

The corresponding normalizing constant is

$$W_i^{mfloc} \;=\; \sum_{z_i} \exp(-H_i^{mfloc}(z_i)).$$

The mean field theory suggests that the marginal distribution of the field at site i,

$$P_G(z_i) \;=\; W^{-1} \sum_{\mathbf{z}_{S\backslash\{i\}}} \exp(-H(\mathbf{z})),$$

can be approximated by

$$\begin{aligned} P_i^{mf}(z_i) &= {W_i^{mf}}^{-1} \exp(-H_i^{mf}(z_i)) \\ &= {W_i^{mfloc}}^{-1} \exp(-H_i^{mfloc}(z_i)), \end{aligned}$$

which is also the conditional probability of Z_i given $\mathbf{Z}_{N(i)} = \mathbf{m}_{N(i)}$,

$$P_i^{mf}(z_i) \quad = \quad P_G(z_i \mid \mathbf{m}_{N(i)}).$$

For Markov random fields, there is no need to fix other sites than the neighbors (see Appendix of [16] as an illustration).

The mean field approximation of the joint distribution $P_G(\mathbf{z})$ is then given by the product

$$P^{mf}(\mathbf{z}) \quad = \quad \prod_{i \in S} P_i^{mf}(z_i). \tag{14}$$

The main difference with the pseudo-likelihood lies in the fact that in (5) the neighbors are still allowed to fluctuate while in the mean field approximation, they are set to constants making each term in the product (14) independent and P^{mf} a valid probability distribution. In terms of computation, to use the mean field approximation, one needs the mean values at sites j different from i. However, these mean values are unknown and it is actually the goal of the approximation to compute them. Therefore, the method depends on a self-consistency condition which is that the mean computed based on the approximation must be equal to the mean used to define this approximation. Then, replace in our previous notation, the exact mean values $m_j, j \in S$ by the mean values in the approximation, denoted by $\bar{z}_j, j \in S$. The same expressions as before hold and we shall not modify our notation. For example, we shall write $H_i^{mf}(z_i)$ with in mind,

$$H_i^{mf}(z_i) \quad = \quad H(\mathbf{z})|_{z_j = \bar{z}_j, j \neq i} \; = \; H(z_i \bar{\mathbf{z}}_{S \setminus \{i\}}),$$

instead of (12). Let $\mathbb{E}_i^{mf}$ denote the expectation under P_i^{mf}. For pixels j different from i, we clearly have $\mathbb{E}_i^{mf}[Z_j] = \bar{z}_j$, as desired, but for site i, it comes

$$\mathbb{E}_i^{mf}[Z_i] \quad = \quad W_i^{mf^{-1}} \sum_{z_i} z_i \exp(-H_i^{mf}(z_i)) \; = \; W_i^{mfloc^{-1}} \sum_{z_i} z_i \exp(-H_i^{mfloc}(z_i)).$$

The last expression is a function of $\{\bar{z}_j, j \in N(i)\}$ that we shall denote by $g_i(\{\bar{z}_j, j \in N(i)\})$. Applying the self-consistency condition leads to

$$\bar{z}_i = \mathbb{E}_i^{mf}[Z_i] = g_i(\{\bar{z}_j, j \in N(i)\}).$$

Repeating this for all sites gives the equation below (n is the number of sites in S)

$$\bar{\mathbf{z}} = g(\bar{\mathbf{z}}) = \begin{cases} g_1(\{\bar{z}_j, j \in N(1)\}) \\ \vdots \\ g_n(\{\bar{z}_j, j \in N(n)\}) \end{cases} . \tag{15}$$

The mean field approximation consists of solving this fixed point equation and taking the solution $\bar{\mathbf{z}} = \{\bar{z}_i, i \in S\}$ as an estimate of the exact mean field $\mathbf{m}$. Equation (15) can also be recovered from a different point of view, namely from the minimization, over the set of probability distributions P that factorize, of quantity $\mathbb{E}_P[\log(\frac{P(\mathbf{Z})}{P_G(\mathbf{Z})})]$ which is the Kullback-Leibler divergence between P and the true Gibbs distribution P_G, (see [5] for more detail). It shows that the mean field approximation is optimal (in the sense of the Kullback-Leibler divergence) among systems of independent variables. Existence and uniqueness of solutions of (15) will not be discussed here but when it exists such a solution is usually computed iteratively (see [17] and [18]).

4 Mean Field-like approximations

The mean field approach consists of neglecting fluctuations from the mean in the environment of each pixel. More generally, we talk about mean field-like approximations when the value at site i does not depend on the values at other sites which are all set to constants (not necessarily the means) independently of the value at site i. We apply this idea to release the computational burden when dealing with the complex joint distribution $P_G(\mathbf{y}, \mathbf{z} \mid \theta, \beta)$ in the EM procedure described in Section 2.3. It follows a class of algorithms described in the next section.

4.1 EM algorithm-based procedures

The general form of the algorithms that we propose consists of repeating the following steps,

(1) Create, from the observations $\mathbf{y}$ and some current parameter estimates $\Psi^{(q-1)}$, a configuration $\tilde{\mathbf{z}}^{(q)}$. For each site i, set the neighbors to $\tilde{\mathbf{z}}^{(q)}_{N(i)}$ and replace the marginal distribution $P_G(\mathbf{z} \mid \beta)$ by

$$P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta) = \prod_{i \in S} P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)}, \beta). \tag{16}$$

(2) Apply the EM algorithm for the model defined by (6) and (16), with starting values $\theta^{(q-1)}$ and $\beta^{(q-1)}$, to get updated estimates $\theta^{(q)}$ and $\beta^{(q)}$. The joint distribution $P_G(\mathbf{y}, \mathbf{z} \mid \Psi)$ is thus replaced by

$$\prod_{i \in S} \{ f_i(y_i \mid z_i, \theta) \; P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)}, \beta) \}, \tag{17}$$

which corresponds to an observed likelihood of the form

$$\begin{aligned} P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{y} \mid \Psi) &= \sum_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{z}, \theta) P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta) \\ &= \prod_{i \in S} \sum_{z_i} f_i(y_i \mid z_i, \theta) \; P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)}, \beta) \\ &= \prod_{i \in S} P_G(y_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)}, \Psi). \end{aligned} \tag{18}$$

This general procedure using (18) has been proposed by [19]. They called their estimation algorithm based on (18) the point-pseudo-likelihood (PPL)-EM algorithm. Because of step (1), the two problems encountered when considering the EM algorithm with the exact joint distribution disappear. The computation of the normalizing constant in $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta)$ becomes easy, and the E step reduces to the computation, in (8), of conditional probabilities corresponding to the approximation of the conditional distribution $P_G(\mathbf{z} \mid \mathbf{y}, \Psi^{(q)})$. This approximation derives naturally from the approximation (16) of $P_G(\mathbf{z} \mid \beta)$ as

$$P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(q)}) = \frac{f(\mathbf{y} \mid \mathbf{z}, \theta^{(q)}) \; P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \beta^{(q)})}{P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{y} \mid \Psi^{(q)})}, \tag{19}$$

which can be further simplified, using (6) and (18), into

$$\begin{aligned} P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(q)}) &= \prod_{i \in S} \{ \frac{f_i(y_i \mid z_i, \theta^{(q)}) \; P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)}, \beta^{(q)})}{\sum_{z_i} f_i(y_i \mid z_i, \theta^{(q)}) \; P_G(z_i \mid \tilde{\mathbf{z}}^{(q)}_{N(i)}, \beta^{(q)})} \} \\ &= \prod_{i \in S} P_G(z_i \mid y_i, \tilde{\mathbf{z}}^{(q)}_{N(i)}, \Psi^{(q)}) \\ &= \prod_{i \in S} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid y_i, \Psi^{(q)}). \end{aligned} \tag{20}$$

It follows the corresponding approximations of $Q(\theta \mid \Psi^{(q)})$ and $Q(\beta \mid \Psi^{(q)})$, defined in (9) and (10),

$$Q(\theta \mid \Psi^{(q)}) \approx \sum_{i \in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid y_i, \Psi^{(q)}) \; \log f_i(y_i \mid z_i, \theta),$$

and

$$Q(\beta \mid \Psi^{(q)}) \approx \sum_{i \in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid y_i, \Psi^{(q)}) \, \log P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \beta).$$

The following M step in EM becomes tractable. As noted by [19], the likelihood (18) takes the form of a likelihood from independent observations from finite mixture of the same component densities but the sets of mixing weights vary for each site i depending on the choice of $\tilde{\mathbf{z}}^{(q)}$.

Our approach can be seen as an attempt to make the EM algorithm for hidden Markov fields more tractable by approximating the complete likelihood $P_G(\mathbf{y}, \mathbf{z} \mid \Psi)$. We derive consistent approximations, in the sense that the Bayes rule is satisfied, of quantities like $P_G(\mathbf{z} \mid \beta)$ and $P_G(\mathbf{y} \mid \Psi)$ (equations (19) and (18)). Our procedures can be seen as the EM algorithm applyed with a different independent mixture model at each iteration, corresponding to the simplified distribution (16). This is not necessary the case in procedures such as the one presented in Section 2.3 [4] combining pseudo-likelihood approximation and MCMC simulations. In this case, the approximation method does not lead to a single valid model. Moreover, the spirit of our general procedure is to consider $\tilde{\mathbf{z}}^{(q)}$ in step (1), as a set of values used to approximate the neighbors interactions rather than as a possible current restoration. In the mean field approximation, for instance, $\tilde{\mathbf{z}}^{(q)}$, whose components are not necessarily discrete values, may not even be a valid configuration for the Markov field.

4.2 Choosing the neighbors

The flexibility of our procedure is then in the choice of the values $\tilde{\mathbf{z}}^{(q)}$. A natural candidate would be one that leads to a reasonable approximation of $P_G(\mathbf{y}, \mathbf{z} \mid \theta, \beta)$. In our model, $P_G(\mathbf{z} \mid \beta)$ and $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$ are not available while $f(\mathbf{y} \mid \mathbf{z}, \theta)$ is. Knowing $f(\mathbf{y} \mid \mathbf{z}, \theta)$, it is enough to approximate one of the unknown quantities, either $P_G(\mathbf{z} \mid \beta)$ or $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$, to derive an approximation of the other and of the joint distribution (equations (19) and (17)). Therefore, our selection of $\tilde{\mathbf{z}}^{(q)}$ can be driven by the quality of the corresponding approximation of $P_G(\mathbf{z} \mid \beta)$ or $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$. As regards the Kullback-Leibler divergence, the approximations cannot be both optimal and satisfy the Bayes rule (19). It seems more reasonable to base our choice on the conditional field distribution rather than on the marginal field distribution. It has the advantage of taking the observations directly into account. Moreover, the study of the case of the homogeneous isotropic Potts model gives reasons disuading from using the mean field approximation on the marginal field (see Appendix of [16] and [20]). Note also that when β is known and need not to be estimated, only $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$ is needed. Hereunder, we consider different ways of approximating the conditional distribution. A reference choice is the mean field approximation [6], [7], [21], where $\tilde{\mathbf{z}}^{(q)}$ is the mean field approximation of the mean of the conditional distribution, with the unknown Ψ replaced by the current estimate $\Psi^{(q)}$. This approximation of the conditional distribution induces, by Bayes theorem, a natural approximation of the marginal distribution, which is not the mean field approximation of the marginal distribution.

A generalization of the conditional mean field approximation is to set the neighbors to constants not necessarily equal to the mean values. For instance, the neighbors can be set to a mode (if many possible) of their conditional distribution. These modes are unknown but we can use a self-consistency condition, as in the mean field theory, saying that the mode computed based on the approximation must have the same value as the mode used to define the approximation. Let z_i^* be a mode of the approximate conditional distribution at site i. As before, $P_G(z_i \mid \beta)$ is approximated by

$$P_i^{mode}(z_i \mid \beta) = P_G(z_i \mid \mathbf{z}^*_{N(i)}, \beta),$$

and therefore, $P_G(z_i \mid \mathbf{y}, \Psi)$ is approximated by $P_G(z_i \mid \mathbf{y}, \mathbf{z}^*_{N(i)}, \Psi)$ so that the associated self-consistency condition is

$$z_i^* \;=\; \arg\max_{z_i} P_G(z_i \mid \mathbf{y}, \mathbf{z}^*_{N(i)}, \Psi).$$

Repeating this for all i in S leads to a fixed point equation that we can solve iteratively. This actually leads to the Iterated Conditional Mode (ICM) algorithm of [2] when the parameters are known.

Another choice that we considered for step (1), consists of simulating $\tilde{\mathbf{z}}^{(q)}$ as a realization of the conditional distribution using the Gibbs sampler, as presented in [1] with the temperature parameter T set to 1. The implementation detail is given in Section 4.3. Those three choices lead to three algorithms that we will refer to respectively as the *mean field*, *mode field* and *simulated field* algorithms. To summarize, step (1) consists of

mean field algorithm:

setting $\tilde{\mathbf{z}}^{(q)}$ to the mean field estimate of the conditional distribution $P_G(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$;

mode field algorithm:

setting $\tilde{\mathbf{z}}^{(q)}$ to the mode field estimate of the conditional distribution $P_G(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$;

simulated field algorithm:

simulating $\tilde{\mathbf{z}}^{(q)}$ from the conditional distribution $P_G(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$.

Ideally, one would like to work with the *best* approximation among systems of independent variables. When Ψ is known and not to be estimated, the procedure described in Section 4.1 reduces to step (1). The three choices above correspond to optimal solutions in three different ways. In the mean field case, the distribution $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y}, \Psi)$ with $\tilde{\mathbf{z}}^{(q)}$ solution of the corresponding fixed point equation is the best approximation of $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$ in the sense of the Kullback-Leibler divergence. For the mode field algorithm, $\tilde{\mathbf{z}}^{(q)}$ converges towards a local maximum of $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$, but it is not clear how well $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y}, \Psi)$ approximates the true conditional distribution. In the simulated field algorithm, step (1) is solved by running a Gibbs sampler so that at convergence $\tilde{\mathbf{z}}^{(q)}$ is a realization of the true conditional distribution $P_G(\mathbf{z} \mid \mathbf{y}, \Psi)$.

4.3 Implementation

In this section, we specify how we implement our algorithms in practice. In principle, in the mean field, mode field and simulated field algorithms, step (1) consists respectively of solving a fixed point equation such as (15), carrying out an ICM algorithm (see [2]) with parameters set to the current value $\Psi^{(q-1)}$, and simulating from a fixed distribution. Each of these three versions involves an iterative procedure in which the updating of $\tilde{\mathbf{z}}^{(q)}$ can be proceeded in two different ways: synchronously or sequentially. In the latter case, each site is updated in turn, using the new values of the other sites as soon as they become available rather than waiting until all sites have been updated. In practice, [10] noticed that solving the fixed point equation in the mean field algorithm, using a sequential updating, leads to faster convergence and avoids possible oscillations. For ICM also (see [2] and [17]), the convergence is guaranteed in the sequential case but not in the synchronous one. In the simulated field algorithm, a synchronous version of the Gibbs sampler is invalid [1]. A stationary distribution will still exits, but its nature is not identified. For these experimental and theoretical reasons, the updating is done sequentially in our algorithms, reporting each new value $\tilde{z}_i^{(q)}$ before addressing the other sites. This corresponds to a global strategy we adopted for our algorithms in which each new value in the set $\{\mathbf{z}, \theta, \beta\}$ is immediately taken into account when updating the others. In particular, we chose to run only one iteration of the procedures defining step (1). As in [2], only one *cycle* is carried out, using $\tilde{\mathbf{z}}^{(q-1)}$ and the current estimate $\Psi^{(q-1)}$, a *cycle* corresponding to the updating of all n pixels. Also, we used a single iteration of EM in step (2). (In their PPL-EM algorithm, [19] used a few iterations of EM.) Because of these choices, the algorithms we actually used in our experiments are the ones detailed below. Let $\tilde{\mathbf{z}}^{(q-1+\frac{i}{n})}$ be $(\tilde{\mathbf{z}}_1^{(q)}, \ldots, \tilde{\mathbf{z}}_i^{(q)}, \tilde{\mathbf{z}}_{i+1}^{(q-1)}, \ldots, \tilde{\mathbf{z}}_n^{(q-1)})$, the configuration updated until site i. One iteration of the procedures defining step (1), using a sequential updating, consists respectively of,

(1) *Choosing the neighbors*
mean field algorithm:

$$\tilde{\mathbf{z}}^{(q)} = I\!E_{P_{\tilde{\mathbf{z}}^{(q-1)}}(\,.\mid \mathbf{y},\Psi^{(q-1)})}[\mathbf{Z}]$$

which according to (20) is equivalent, in our sequential version, to

$$\forall i \in S, \quad \tilde{z}_i^{(q)} = I\!E_{P_G(\,.\mid y_i, \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, \Psi^{(q-1)})}[Z_i],$$

$$i.e.\ \ \tilde{z}_i^{(q)} = \frac{\sum\limits_{z_i} z_i \exp\{-\sum\limits_{c\ni i} V_c(z_i, \tilde{\mathbf{z}}_{c\setminus\{i\}}^{(q-1+\frac{i-1}{n})} \mid \beta^{(q-1)}) + \log f_i(y_i \mid z_i,\ \theta^{(q-1)})\}}{\sum\limits_{z_i} \exp\{-\sum\limits_{c\ni i} V_c(z_i, \tilde{\mathbf{z}}_{c\setminus\{i\}}^{(q-1+\frac{i-1}{n})} \mid \beta^{(q-1)}) + \log f_i(y_i \mid z_i,\ \theta^{(q-1)})\}}\ ;$$

mode field algorithm:

$$\tilde{\mathbf{z}}^{(q)} = \arg\max_{\mathbf{z}} P_{\tilde{\mathbf{z}}^{(q-1)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)}),$$

which according to (20) leads, in a sequential form, to

$$\forall i \in S, \quad \tilde{z}_i^{(q)} = \arg\max_{z_i} P_G(z_i \mid y_i, \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, \Psi^{(q-1)})$$

$$i.e.\ \ \tilde{\mathbf{z}}_i^{(q)} = \arg\max_{z_i} f_i(y_i \mid z_i, \theta^{(q-1)}) P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, \beta^{(q-1)})\ ;$$

simulated field algorithm:

for all i in S, $\tilde{z}_i^{(q)}$ is simulated from $P_G(z_i \mid y_i, \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, \Psi^{(q-1)})$,
which is proportional to $f_i(y_i \mid z_i, \theta^{(q-1)}) P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, \beta^{(q-1)})$.

We then run a single iteration of the EM algorithm using the distribution $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{y}, \mathbf{z} \mid \Psi)$ instead of $P_G(\mathbf{y}, \mathbf{z} \mid \Psi)$, so that the second step is,

(2) *EM iteration*

(E) compute $P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \mathbf{y}, \Psi^{(q-1)})$ for all i in S;

(M) set $\Psi^{(q)} = (\theta^{(q)}, \beta^{(q)})$ with

$$\theta^{(q)} = \arg\max_{\theta} \sum_{i\in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \mathbf{y}, \Psi^{(q-1)})\ \log f_i(y_i \mid z_i, \theta),$$

and

$$\beta^{(q)} = \arg\max_{\beta} \sum_{i \in S} \sum_{z_i} P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \mathbf{y}, \Psi^{(q-1)}) \; \log P_{\tilde{\mathbf{z}}^{(q)}}(z_i \mid \beta). \tag{21}$$

5 Related algorithms

In place of the EM algorithm, other algorithms can be considered in step (2), as the Classification EM (CEM) algorithm [22] or the Stochastic EM (SEM) algorithm [23]. They both consist of generating a configuration $\mathbf{z}^{(q)}$ after the E step and use it as an image restoration in the following M step. In the CEM algorithm, $\mathbf{z}^{(q)}$ is generated according to a maximum a posteriori (MAP) rule (C step) while it is simulated in the SEM algorithm (S step). In our sequential implementation of a single iteration of CEM or SEM, step (2) turns as follows,

(E) same E step than in EM.

The additional step is given by

(C) for CEM,

$$\forall i \in S, \quad z_i^{(q)} = \arg\max_{z_i} P_G(z_i \mid y_i, \tilde{\mathbf{z}}_{N(i)}^{(q)}, \Psi^{(q-1)}),$$

(S) for SEM,

for all i in S, $z_i^{(q)}$ is simulated from $P_G(z_i \mid y_i, \tilde{\mathbf{z}}_{N(i)}^{(q)}, \Psi^{(q-1)})$, which is proportional to $f_i(y_i \mid z_i, \theta^{(q-1)}) P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q)}, \beta^{(q-1)})$.

Considering $\mathbf{z}^{(q)}$ as an image restoration, the M step becomes

(M) set $\Psi^{(q)} = (\theta^{(q)}, \beta^{(q)})$ with

$$\theta^{(q)} = \arg\max_{\theta} \sum_{i \in S} \log f_i(y_i \mid z_i^{(q)}, \theta),$$

and

$$\beta^{(q)} = \arg\max_{\beta} \sum_{i \in S} \log P_{\tilde{\mathbf{z}}^{(q)}}(z_i^{(q)} \mid \beta). \tag{22}$$

More generally, any combination of ways to generate the neighbors in step (1) and algorithms in step (2) can be considered. However, some combinations may appear more natural. For instance, the mode field approximation combined with CEM is almost the unsupervised ICM of [2]. It is recovered exactly if, in the E step, $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(q-1)})$ is replaced by a degenerate distribution giving probability one to configuration $\tilde{\mathbf{z}}^{(q)}$. This is equivalent to set $\mathbf{z}^{(q)} = \tilde{\mathbf{z}}^{(q)}$, so that (22) consists of maximizing the pseudo-likelihood. Reading already known and commonly used procedures in our general framework is interesting. It brings a new and insightful view on these algorithms. As an illustration, if only one EM iteration is performed in step (2), the *mean field algorithm* is the algorithm of [6] with the addition that our approach provides a reason for choosing the approximation of the marginal Markovian distribution as it is proposed with no justification in [6]. In our framework, this choice is consistent with the preferable and recommended approximation of the conditional Markovian distribution. Also, distribution (18) has the same form than the Point-Pseudo-Likelihood (PPL) used by [19], so that their PPL-EM algorithms are included in our class. Note that the authors suggest performing step (1) using the MAP or ICM. The algorithm of [6] and the apparently distant PPL-EM algorithms appear to arise from the same approximation principle.

An important feature in our mean field-like algorithms is that, as opposed to ICM-like algorithms, we work with probabilities rather than with classifications based on a Maximum A Posteriori (MAP) rule. The current $\tilde{\mathbf{z}}^{(q)}$ is not considered as a current hidden field. Only approximations of the marginal and the conditional distributions are used to estimate the parameter Ψ, even if a configuration $\tilde{\mathbf{z}}^{(q)}$ is needed to define these approximations. These procedures actually treat the $\tilde{\mathbf{z}}^{(q)}_{N(i)}$'s as if they were the truth but the z_i's are still assumed to be unknown. This appears to lead to less biased estimates (see Section 6.1 for an illustration).

6 Experiments

We experimented on simple models, using 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,

$$P_G(\mathbf{z} \mid \beta) \quad = \quad W(\beta)^{-1} \exp(\beta \sum_{i \sim j} z_i^t z_j) \;, \tag{23}$$

where the notation $i \sim j$ represents all couples of sites (i, j) which are neighbors. In that case, step (1) of the algorithms described in Section 4.1 can be further specified. The mean field approximation applies provided appropriate formulations of the clique potentials or energy function. The conditional mean field fixed point iteration becomes, in our sequential form,

$$\forall i \in S \;, \quad \tilde{z}_i^{(q)} = \frac{\sum\limits_{z_i} z_i \exp\{\beta^{(q-1)} z_i^t \sum\limits_{j \in N(i)} \tilde{z}_j^{(q-1+\frac{i-1}{n})} + \log f_i(y_i \mid z_i,\ \theta^{(q-1)})\}}{\sum\limits_{z_i} \exp\{\beta^{(q-1)} z_i^t \sum\limits_{j \in N(i)} \tilde{z}_j^{(q-1+\frac{i-1}{n})} + \log f_i(y_i \mid z_i,\ \theta^{(q-1)})\}} \;,$$

and the quantity used in the mode and simulated field algorithms is proportional to,

$$f_i(y_i \mid z_i, \theta^{(q-1)}) P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}^{(q-1+\frac{i-1}{n})}, \beta^{(q-1)}) \;\propto\; f_i(y_i \mid z_i, \theta^{(q-1)}) \exp(\beta^{(q-1)} z_i^t \sum_{j \in N(i)} \tilde{z}_j^{(q-1+\frac{i-1}{n})}) \;.$$

See Appendix of [16] for more detail on the mean field approximation of the Potts model.

For the f_i's we considered Gaussian distributions. 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 scalar mean and standard deviation values, in the univariate situation, and vector means and covariance matrices in the multivariate case. The parameter to be estimated is then $\{\beta, \theta\}$ with $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$.

As mentioned in Section 4.3, we used sequential updatings and carried out only one cycle in step (1). We noticed no significant improvement by running more cycles.

In this section, we report some experiments on the three algorithms described in this article. For comparison we also provide the results when applying the EM algorithm for independent mixture model, and when using ICM. The EM algorithm for independent mixtures, denoted by EM hereafter, is the only algorithm among the ones we tested that does not take into account spatial information. We considered an unsupervised version of ICM where parameter β is estimated at each iteration by maximizing the current pseudo-likelihood. The algorithms were all initialized using the same segmentation obtained by simple thresholding: We divided the pixel value range, in the degraded image, into regular intervals and assigned each of them to a class.

For the mean, mode and simulated field algorithms, the restorations shown result from the maximization of the conditional distribution $P_{\tilde{\mathbf{Z}}^{(N)}}(\mathbf{z} \mid \mathbf{y}, \Psi^{(N)})$ provided by the algorithms. In practice, we set N to 100, observing no significant improvement, for the examples we considered, when carrying out more iterations. The EM algorithm was stopped after 100 iterations and ICM was run until convergence. Note that no significant time differences were observed when running one iteration of the mean, mode, simulated field algorithms and the ICM algorithm.

The code we made to carry out these experiments was initially based on an existing code for spatial classification made by [10] and available at *http://www.hds.utc.fr/ mdang/*.

We investigated four examples. We first compared the algorithms in terms of parameter estimation on simulated data (Section 6.1). In the other examples, we studied the performance in terms of the quality of the segmentations on noise-corrupted images (Section 6.2). Additional related experiments are given to account for various observations on the algorithms behavior (Section 6.3).

6.1 Hidden 2-color Potts models

We first tested the algorithms on images simulated from hidden Potts models for which the true parameters β and θ were known. We created 150×150 binary images by simulating (algorithm of [24]) 2D 2-color Potts models (23) and then adding a Gaussian noise. We considered a first order neighborhood, *i.e.* four neighbors for each pixel. The simulated data correspond to hidden 2-color Potts models for which $\theta = \{(\mu_k, \sigma_k), k = 1, 2\}$ with $\mu_k = k$ and $\sigma_k = 1$, for $k = 1, 2$. We used our knowledge of a constant variance, for the two states, to fit a model and recover the true image. Parameter estimates are given in Table 1 for two different values of β, 0.2 and 0.6. In these cases, the closest estimation of β to the true value is given by the simulated field algorithm while the mean field algorithm overestimates the spatial parameter. Moreover, it appears that ICM overestimates the distance between the two means and underestimates the standard deviation. This is not surprising since ICM lies on a classification obtained according to a MAP principle to estimate the model parameters and consequently tends to produce biased Gaussian parameter estimates (see for instance [25]). The same behavior is observed for the closely related mode field algorithm.

6.2 Noise-corrupted images

In this section we consider three noise-corrupted images corresponding to different number of classes and different noise models. In each case, the degraded image was obtained by adding some Gaussian noise to the K-color original image. For these data, the spatial parameter of the hidden field model is unknown but the true parameters of the simulated noise are available. We used a model with second order neighborhood, (*i.e.* the eight closest neighbors for each pixel), which is in general more realistic.

6.2.1 Four-color test image

The first image is a 128×128 image obtained by adding some Gaussian noise to the 4-color image (a) of Figure 1, leading to Figure 1 (b). 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 $k = 1, \ldots, 4$. In order to test the procedures ability to recover that the variances were equal, we tried to fit a model with class dependent variances. The algorithms were applied to obtain classifications in Figure 1 (d) to (h). The corresponding estimated parameters are given in Table 2. Since the true image is known we also report, in this table, the error rates (*i.e.* the percentages of misclassified pixels) as an indication of the algorithms ability to restore the truth in addition to a visual assessment. The common variance is always recovered but the algorithms using spatial information clearly outperform EM for independent mixture model in terms of restoration. The mean field and simulated field algorithms provide similar estimations of parameter β, which was not the case for the previous synthetic images we analysed. Moreover, they provide better noise parameters estimation and produce notably lower error rates than ICM. The same is true for the mode field algorithm to a lesser extent.

6.2.2 The Logo image

The second example is a 2-color logo image degraded with Gaussian noise. We used Gaussian densities with class-dependent variances so that the true noise parameters are $(\mu_1, \sigma_1) = (51, 130)$ and $(\mu_2, \sigma_2) = (255, 300)$. The segmentations obtained with the different methods are shown in Figure 2 (d) to (h) with the corresponding estimated parameters in Table 3. As in the previous example, the algorithms using spatial information clearly outperform the EM algorithm for independent mixture models that keeps a lot of noisy pixels. Among the spatial procedures, the mean field and simulated field algorithms give comparable results and perform better than ICM and the mode field algorithm.

6.2.3 The Mickey mouse image

The particularity of this image (Figure3 (a)) is the presence of thin lines representing the cheeks, smile and eyes. The segmentations of the observed image 3 (b) after applying the different algorithms are given in Figure 3 (d) to (h). This case is interesting in that it illustrates the possible oversmoothing behavior of the mean field algorithm (Figure 3 (e)) while the other spatial algorithms are still able to recover most of the thin lines. In particular, the simulated algorithm that gave similar results than the mean field algorithm when this one was performing well, is still providing good results in this case. The same oversmoothing behavior of the mean field algorithm is observed on other examples (see Section 6.3).

6.3 Additional experiments

In this section, some additional worth noting behaviors of the algorithms are illustrated.

The oversmoothing behavior of the mean field algorithm is observed on other examples. An interesting case is the one illustrated in Figure 4. In this case, it seems to be related to a slower convergence of the mean field algorithm that we observed in several other experiments not reported here. The window frame, for example, is recovered in 100 iterations for the simulated field algorithm while it takes 300 iterations to recover a similar segmentation with the mean field algorithm. Another example is that of Figure 2 (a) corrupted with a larger noise ($\sigma_1 = \sigma_2 = 300$). It is reported in

Figure 5. The difference with the mean and simulated field algorithms after 500 iterations (Figures 5 (c) and (d)) is even more dramatic. The simulated field algorithm was also runned for 1000 and 1500 iterations leading to similar segmentations. However this case also illustrates that, when images generated from severely overlapping mixtures are considered the simulated field performance decreases as the noise increases (Figure 5 (c)). This occurs to a lesser extent for ICM since in this particular situation the use of MAP classifications turns out to be an advantage (Figure 5 (e)).

7 Discussion

In the context of Markov model-based image segmentation, we have introduced a general framework for deriving procedures based on approximations of dependence between the hidden variables defining the data. These approximations can be seen as a generalization of the mean field approximation principle of statistical physics. They consist of using systems of independent variables by fixing each pixel's neighbors to constants, neglecting the fluctuations of the pixels interacting with it. We have shown the interest of such approximations to deal with the computational complexity and intractability of the EM algorithm in problems where spatial information is important. We focused more specifically on three procedures for which we reported some experiments results. The *mean field* algorithm uses the standard mean field approximation. If in our implementation, we restrict to a single iteration in step (2), this is the procedure proposed by [6] which as a consequence gains better justification when being restated in our framework. The two other procedures are new algorithms. The *mode field* algorithm is closer to schemes like the ICM algorithm of [2], while the *simulated field* algorithm is related to the Gibbs sampler of [1]. Comparing, on synthetic and real images, with the EM algorithm for independent mixture models, the three algorithms show significant improvement in terms of the segmentation smoothness. It confirms the gain in dealing with spatial dependences through approximations by systems of independent variables: the computational simplicity of EM for independent mixtures is preserved while spatial information is included. In particular, our experiments point out the good performance of the new simulated field algorithm. On synthetic examples, this procedure provides the best spatial parameter estimations. As regards segmentations it outperforms or is equivalent to the mean field algorithm and does not seem to encounter the same oversmoothing and slow convergence problems. An interesting feature of both mean field and simulated field algorithms is that they involve probabilities instead of MAP classifications and as a result lead in most cases to more satisfying estimations and segmentations than ICM. Also, the relationship between the simulated field algorithm and the Gibbs sampler and Monte Carlo EM-like algorithms [26] and [27] suggests similar behavior in a much smaller computing time. Therefore, the simulated field algorithm can be considered as a powerful algorithm providing little biased estimates of the HMRF parameters in a quite reasonable computing time.

Possible extension of our work is the use of other models for the unobserved image. The basic Potts models (no external field), although performing well in a lot of cases, may not always capture the image characteristics well enough. A natural generalization is to use Potts models with potentials on singletons (site-dependent external field). This induces additional parameter estimation problems but should result in more flexibility and better adequacy when the colors proportions are very unbalanced in the images to be recovered. Another approach is to consider more complex Markov models proposed and studied by [28] and [29]. They are higher order interaction Markov random fields which involve three parameters regulating the presence of noise, edge and line configurations. Our study can be easily extended to these models. Approximations are then all the more interesting as these models require more computation than the simple Potts models.

References

[1] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

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

[3] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

[4] B. Chalmond. An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, 22(6):747–761, 1989.

[5] D. Chandler. *Introduction to Modern Statistical Mechanics.* Oxford University Press, 1987.

[6] J. Zhang. The Mean Field Theory in EM Procedures for Markov Random Fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583, 1992.

[7] J. Zhang. The Mean Field Theory in EM Procedures for Blind Markov Random Field Image Restoration. *IEEE Transactions on Image Processing*, 2(1):27–40, 1993.

[8] G. J. McLachlan and D. Peel. *Finite Mixture Models.* Wiley, 2000.

[9] A. P. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, series B*, 39:1–38, 1977.

[10] V. M. Dang. *Classification de Données Spatiales: Modèles Probabilistes et Critères de Partionnement.* PhD thesis, Université de Technologie de Compiègne, France, 1998.

[11] D. Geiger and F. Girosi. Parallel and deterministic algorithms from MRFs: Surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):401–412, 1991.

[12] J. Zerubia and R. Chellappa. Mean field approximation using compound Gauss-Markov random field for edge detection and image restoration. In *ICASSP'90*, pages 2193–2196, 1990.

[13] A. L. Yuille. Generalized deformable models, statistical physics and matching problems. *Neural Computation*, 2:1–24, 1990.

[14] T. S. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning in graphical models, Jordan, M.I. (Ed.)*, pages 163–173. Dordrencht, Kluwer Academic Publishers, 1998.

[15] T. Hofmann and J. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1–14, 1997.

[16] G. Celeux, F. Forbes, and N. Peyrard. EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation. Research Report no. 4105, Inria, available at http://www.inria.fr/rrrt/rr-4105.html. 2001.

[17] C-H. Wu and P. C. Doerschuk. Cluster Expansions for the Deterministic Computation of Bayesian Estimators Based on Markov Random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):275–293, 1995.

[18] C. Ambroise and G. Govaert. Convergence proof of an EM-type algorithm for spatial clustering. *Pattern Recognition Letters*, 19:919–927, 1998.

[19] W. Qian and D. M. Titterington. Estimation of parameters in hidden Markov models. *Phil. Trans. R. Soc. Lond. A*, (337):407–428, 1991.

[20] G. E. B. Archer and D. M. Titterington. Parameter estimation for hidden Markov chains. *To appear in Journal of Statistical Planning Inference*, 2000.

[21] A.P. Dunmur and D.M. Titterington. Mean fields and two-dimensional markov random fields in image analysis. *Pattern Analysis and Applications*, 1:248–260, 1998.

[22] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332, 1992.

[23] G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly*, 2(1):73–82, 1985.

[24] R. H. Swendsen and J. S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88, 1987.

[25] D. M. Titterington. Comments on a paper by S. L. Sclove. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):656–658, 1984.

[26] G.C.G. Wei and M. A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

[27] M. L. Comer and E. J. Delp. The EM/MPM algorithm for segmentation of textures images: Analysis and further experimental results. *IEEE Transactions on Image Processing*, 9(10):1731–1744, 2000.

[28] X. Descombes, J-F. Mangin, E. Pechersky, and M. Sigelle. Fine structures preserving Markov model for image processing. In *9th Scandinavian Conference on Image Analysis*, 1995.

[29] H. Tjelmeland and J. Besag. Markov random field with higher-order interactions. *Scand. Journ. Stat.*, 25:415–433, 1998.

About the Author Gilles Celeux is Director of Research at Institut National de Recherche en Informatique et Automatique (INRIA). He obtained his "thèse d'état" in Statistics in 1987 from University Paris 9-Dauphine. Since 1993, he is with INRIA Rhône-Alpes and is Head of IS2 a research team on Statistical Modeling with industrial and biomedical applications. His research interests are in statistical modeling, statistical pattern recognition, hidden structure models and reliability.

About the Author Florence Forbes is Research Scientist at Institut National de Recherche en Informatique et Automatique (INRIA). She gained her Ph.D. in Applied Probabilities in 1996 at University Joseph Fourier, Grenoble, France. She joined IS2 research team at INRIA Rhône-Alpes in 1998. Her research activities include Bayesian Image Analysis, Markov processes, Markov Random Fields and hidden structure models.

About the Author Nathalie Peyrard is currently a member of the VISTA research team, doing a postdoc at Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) in Rennes, France. She obtained her Ph.D. in Statistics in 2001 from University Joseph Fourier, Grenoble. Her research interests include spatial statistics, Markov Random Fields, Stochastic algorithms (MCMC) and applications in Image Analysis.

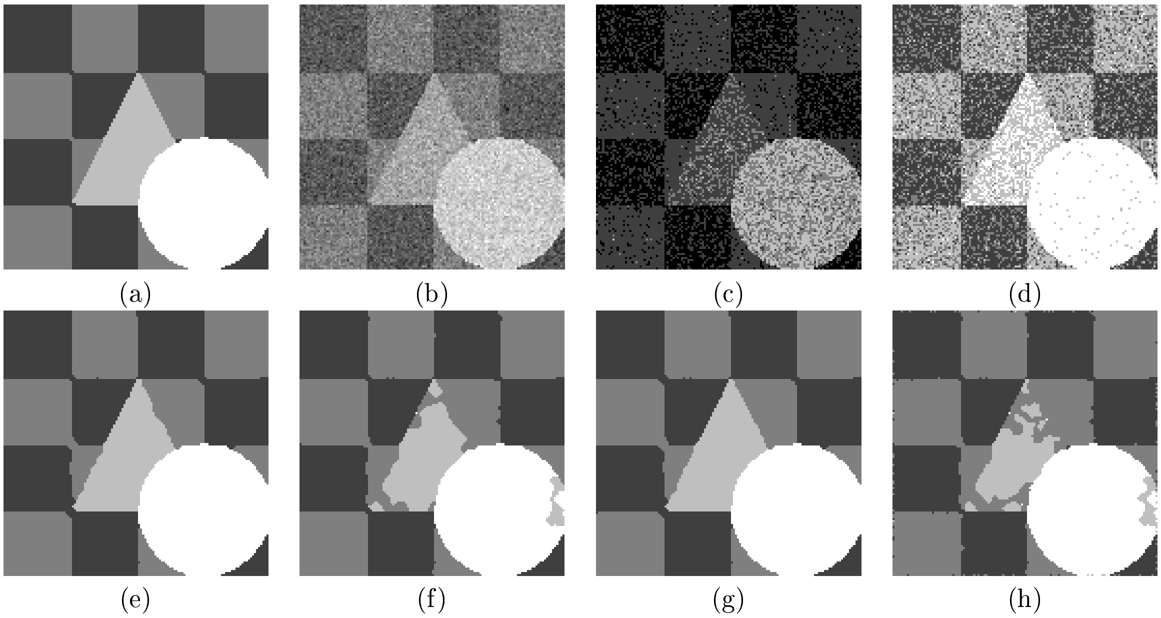


Figure 1:

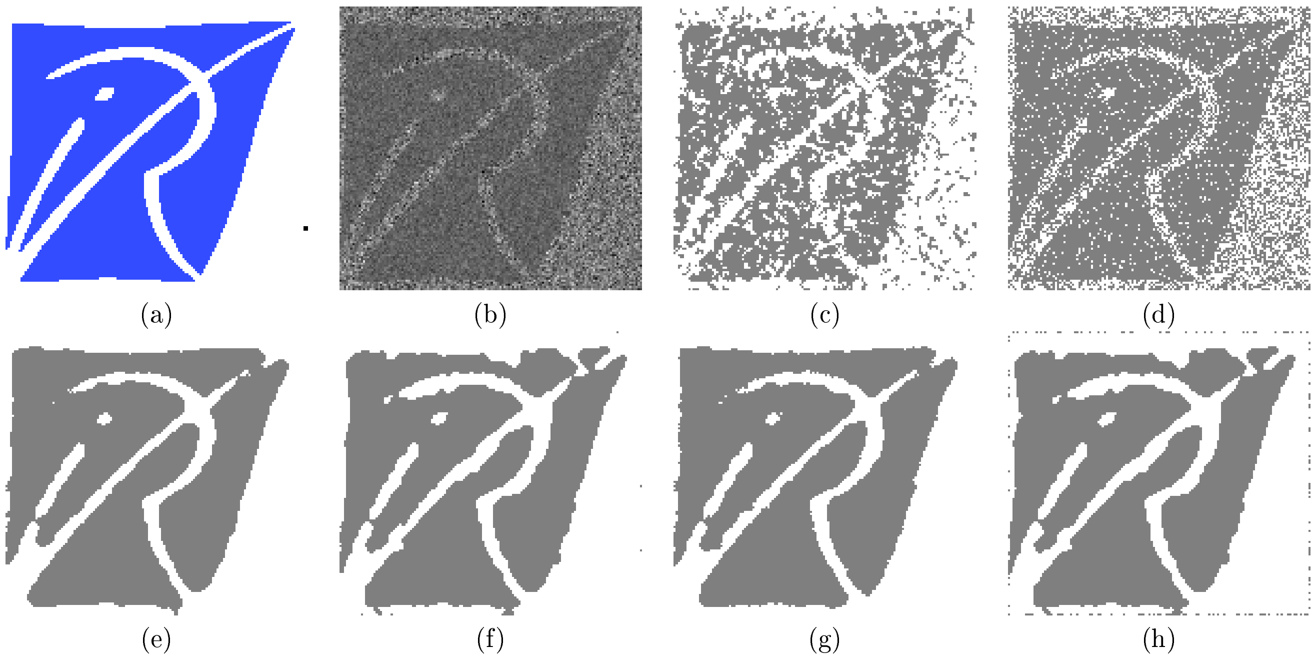

(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 2:

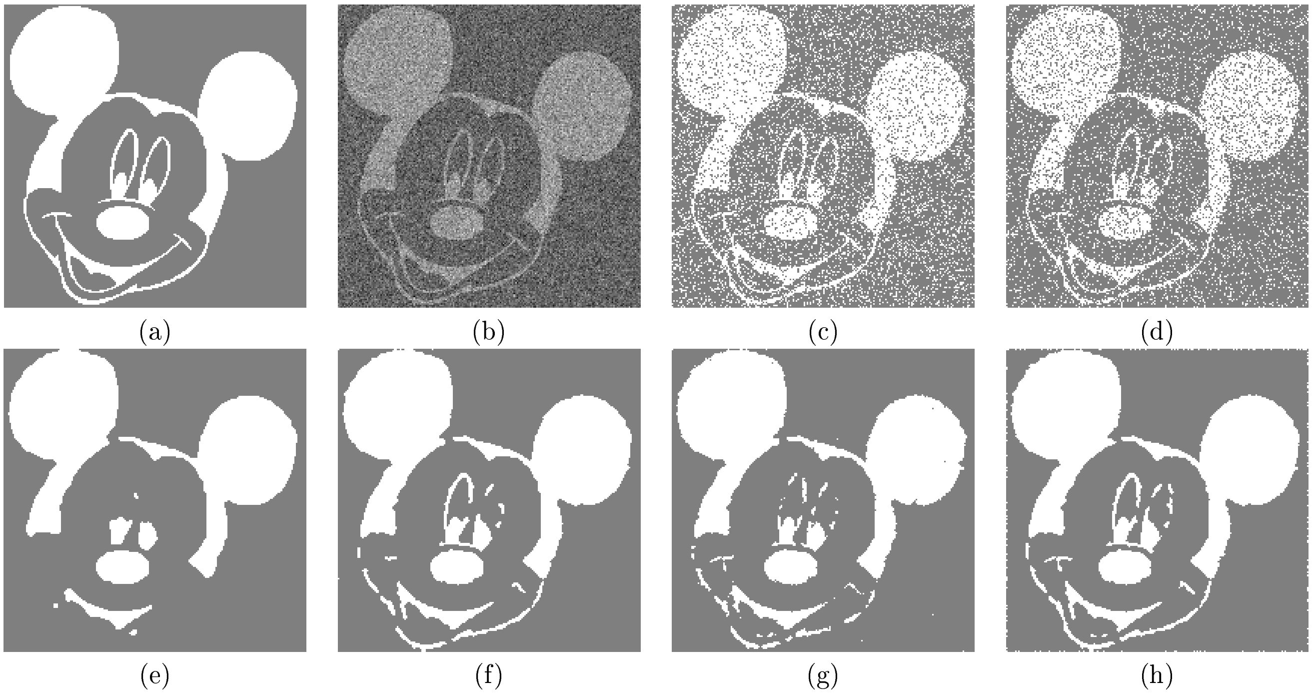
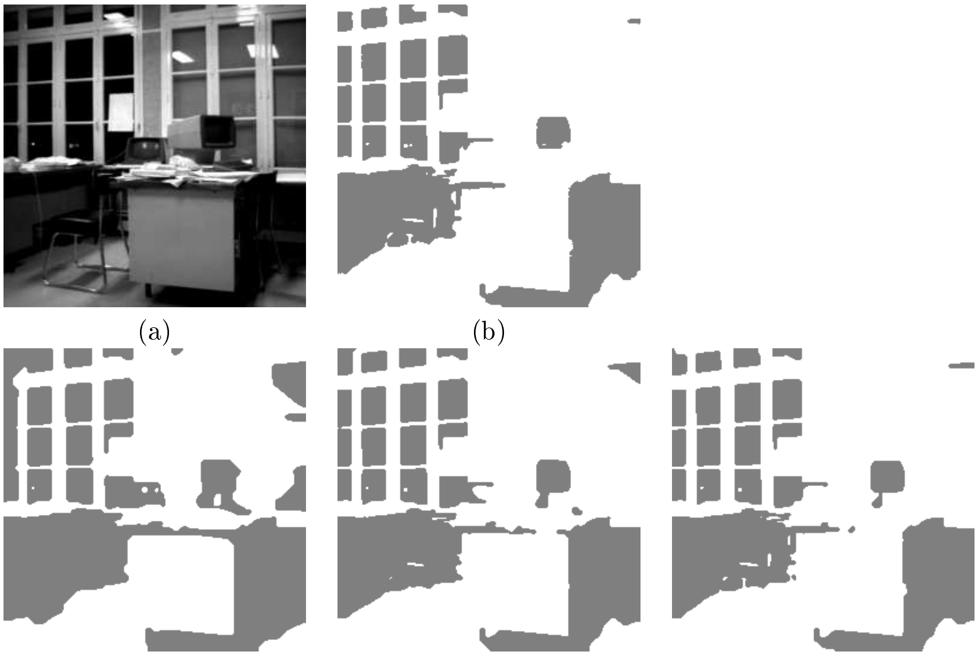


Figure 3:



(a) (b)

(c) (d) (e)

Figure 4:

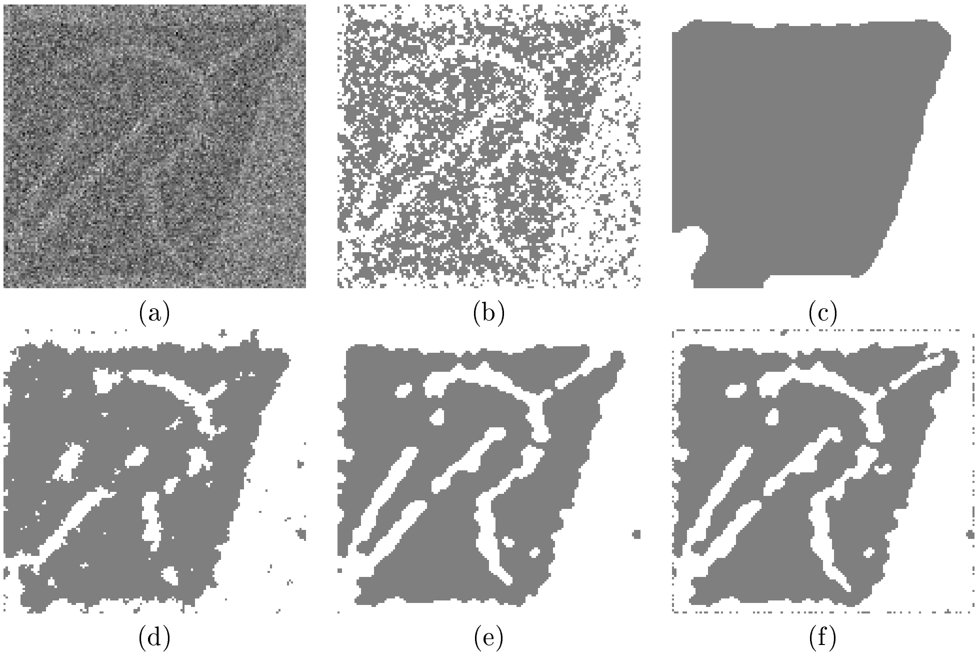


Figure 5:

	β	μ_1	μ_2	σ
true values	0.2	1	2	1
EM	-	1.03	2.05	1.00
Mean Field	1.02	1.26	1.71	1.10
Mode Field	0.52	0.87	2.16	0.92
Simulated Field	0.14	0.99	2.02	1.00
ICM	0.10	0.63	2.42	0.67
true values	0.6	1	2	1
EM	-	0.98	1.97	1.00
Mean Field	1.40	1.25	1.77	1.09
Mode Field	0.85	0.90	2.08	0.95
Simulated Field	0.54	0.97	2.02	1.00
ICM	0.28	0.60	2.37	0.68

Table 1:

	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	-
EM	-	0.8	1.8	3.5	4.1	0.4	0.7	0.7	0.4	27.4
Mean Field	2.0	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5	0.5
Mode Field	1.8	1.0	2.1	3.1	4.0	0.5	0.5	0.5	0.5	2.8
Simulated Field	2.1	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5	0.4
ICM	1.1	1.0	2.1	3.2	4.0	0.5	0.6	0.5	0.5	4.6

Table 2:

	β	μ_1	μ_2	σ_1	σ_2	error rate
true values	-	51	255	130	300	-
EM	-	52	255	128	304	22.11
Mean Field	1.26	52.4	253	129	303	3.24
Simulated Field	0.80	52.4	253	128	304	3.52
Mode Field	1.01	50.2	245	127	299	5.63
ICM	0.89	46.8	251	128	296	5.94

Table 3:

Figure 1: A 4-color image degraded with noise. (a) original image, (b) simulated image, (c) initial segmentation using simple thresholding, (d) EM segmentation, (e) mean field segmentation, (f) mode field segmentation, (g) simulated field segmentation, (h) ICM segmentation.

Figure 2: A 2-color logo image degraded with noise. (a) original image, (b) simulated image, (c) initial segmentation using simple thresholding, (d) EM segmentation, (e) mean field segmentation, (f) mode field segmentation, (g) simulated field segmentation, (h) ICM segmentation .

Figure 3: The Mickey mouse image. (a) original image, (b) simulated image, (c) initial segmentation using simple thresholding, (d) EM segmentation, (e) mean field segmentation, (f) mode field segmentation, (g) simulated field segmentation, (h) ICM segmentation.

Figure 4: The desk image. (a) original image, (b) 2-color segmentation running 100 iterations of the simulated field algorithm, (c), (d), (e) 2-color segmentations running respectively 100, 200 and 300 iterations of the mean field algorithm.

Figure 5: The logo image degraded with heavy noise. (a) simulated image ($\sigma = 300$), (b) initial segmentation using simple thresholding, (c) mean field segmentation, (d) simulated field iteration, (e) mode field segmentation, (f) ICM segmentation.

Table 1: Parameter estimates for the hidden 2-color Potts model with $\beta = 0.2$ and $\beta = 0.6$ (first order neighborhood).

Table 2: Parameter estimates and error rates for the degraded 4-color image.

Table 3: Parameter estimates and error rates for the degraded 2-color logo image.