EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation

Gilles Celeux, Florence Forbes, Nathalie Peyrard

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

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

Gilles Celeux, Florence Forbes, Nathalie Peyrard

N° 4105

Janvier 2001

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--4105--FR+ENG

INRIA
RHÔNE-ALPES

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

Gilles Celeux, Florence Forbes, Nathalie Peyrard

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet is2

Rapport de recherche n ° 4105 — Janvier 2001 — 24 pages

Abstract: This paper deals with Markov random field model-based image segmentation. This involves parameter estimation in hidden Markov models for which one of the most widely used procedures is the EM algorithm. In practice, difficulties arise due to the dependence structure in the models and approximations are required to make the algorithm tractable. We propose a class of algorithms in which the idea is to deal with systems of independent variables. This corresponds to approximations of the pixels' interactions similar to the mean field approximation. It follows algorithms that have the advantage of taking the Markovian structure into account while preserving the good features of EM. In addition, this class, that includes new and already known procedures, is presented in a unified framework, showing that apparently distant algorithms come from similar approximation principles. We illustrate the algorithms performance on synthetic and real images. These experiments point out the ability of our procedures to take the spatial information into account. Our algorithms often show significant improvement when comparing with the EM algorithm applied with no account of the spatial structure and with the ICM algorithm, based on maximization of the pseudo-likelihood and commonly used in image segmentation.

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

(Résumé : tsvp)

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)

Approximations de type champ moyen pour l'utilisation de l'algorithme EM dans les modèles markoviens de segmentation d'images

Résumé : Cet article traite de l'estimation des paramètres d'un champ de Markov caché pour la segmentation d'images. Dans ce cadre, l'algorithme EM, algorithme de référence pour les modèles à structure cachée, se heurte à des difficultés fortes, pour prendre en compte la dépendance spatiale, exigeant des approximations. Nous proposons une classe d'algorithmes fondés sur l'idée de se ramener à des systèmes de variables indépendantes. Cela conduit à des approximations de l'interaction des pixels analogues à l'approximation en champ moyen. Nous obtenons ainsi des algorithmes qui prennent en compte la dépendance markovienne tout en conservant les bonnes caractéristiques de l'algorithme EM. Ce point de vue offre un cadre unifié pour des algorithmes déjà connus ou nouveaux et permet de montrer que des techniques apparemment bien différentes reposent sur des principes d'approximation analogues. Nous illustrons les performances de ces différents algorithmes sur des images simulées et réelles. Ces expérimentations confirment que nos algorithmes prennent bien en compte les hypothèses du modèle de champ de Markov caché. Ils donnent souvent de meilleurs résultats que l'algorithme EM ne tenant pas compte de la dépendance spatiale et que l'algorithme ICM, algorithme de référence en segmentation d'images, basé sur la maximisation de la pseudo-vraisemblance.

Mots-clé : Segmentation d'images, Champs de Markov cachés, Algorithm EM, Algorithme ICM, Pseudo vraisemblance, Approximation en champ moyen, Algorithme du champ simulé.

1 Introduction

Markov random field models revealed themselves as a powerful tool for image segmentation (Geman and Geman 1984, Besag 1986). 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, parameter estimation in hidden Markov models and image segmentation require approximations. A common technique to deal with the models complexity is the use of the pseudo-likelihood approximation (Besag 1975). A reason is that it induces factorizations and makes the models tractable. However it does not necessarily corresponds to a valid probability model. In this paper, the approach we propose is a generalization of the mean field principle of statistical physics (Chandler 1987). More specifically, we consider approximations of Markov models, with complex dependences, by systems of independent variables. These approximations 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 which includes new and already known procedures. Their performance are studied on synthetic and real images.

The following Section 2 specifies the context of hidden Markov models for image segmentation. The mean field approximation principle is presented in Section 3. Our algorithms are described and analysed in Section 4. In Section 5, we consider other algorithms and give additional theoretical comments. 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. To illustrate the mean field approximation principle, an appendix gives some detail on a simple pairwise interaction case.

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 $\{0,1\}^K$. 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,

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

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

where $\mathbf{z}_{S\setminus\{i\}}$ denotes a realization of the field restricted to $S\setminus\{i\} = \{j \in S, j \neq i\}$ 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

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

where H is the energy function

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

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 Besag (1975) and defined as

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

Each term in the product is easy to compute,

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

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),

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

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.* McLachlan and Peel 2000). 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

$$P_G(\mathbf{y}, \mathbf{z} \mid \theta, \beta) = f(\mathbf{y} \mid \mathbf{z}, \theta)\, P_G(\mathbf{z} \mid \beta)$$

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

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 (Dempster, Laird, and Rubin 1977) is an iterative algorithm aiming at maximizing this log-likelihood by maximizing at iteration q,

$$
Q(\Psi \mid \Psi^{(q)}) \ = \ I\!E_{\Psi^{(q)}}[\log P_G(\mathbf{y}, \mathbf{Z} \mid \Psi) \mid \mathbf{Y} = \mathbf{y}]\ ,
$$

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

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

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

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

$$
\begin{aligned}
Q(\beta \mid \Psi^{(q)}) \ = \ I\!E_{\Psi^{(q)}}[\log P_G(\mathbf{Z} \mid \beta) \mid \mathbf{Y} = \mathbf{y}] \ = \ & - \log W(\beta) \\
& - \sum_c \sum_{\mathbf{Z}_c} V_c(\mathbf{z}_c \mid \beta)\ P_G(\mathbf{z}_c \mid \mathbf{y}, \Psi^{(q)}).
\end{aligned} \tag{10}
$$

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 Chalmond 1989). 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 Zhang 1992 and Dang 1998). 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.* Chandler 1987) 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.* Geiger and Girosi 1991, Zerubia and Chellappa 1990, Yuille 1990), graphical models (*e.g.* Jaakkola and Jordan 1998 and references therein) and other areas (*e.g.* Hofmann and Buhmann 1997). 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 $I\!E_G(Z_j)$, 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 $W_i^{mf} = \sum_{z_i} \exp(-H_i^{mf}(z_i))$ 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, $R_i^{mfloc}(\mathbf{m}_{S\setminus\{i\}})$, that does not depend on z_i, namely

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

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,

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

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 $\mathbf{Z}_{N(i)} = \mathbf{m}_{N(i)}$,

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

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

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

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

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,

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

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

$$I\!E_i^{mf}[Z_i] \;=\; {W_i^{mf}}^{-1} \sum_{z_i} z_i \exp(-H_i^{mf}(z_i)) \;=\; {W_i^{mfloc}}^{-1} \sum_{z_i} z_i \exp(-H_i^{mfloc}(z_i)).$$

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

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

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 Chandler 1987 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 Wu and Doerschuk 1995).

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 Qian and Titterington (1991). 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 Qian and Titterington (1991), 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)). This is not necessary the case in procedures such as the one presented in Section 2.3 (Chalmond 1989) combining pseudo-likelihood approximation and MCMC simulations. This procedure does not correspond 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 and Archer and Titterington 2000). 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 (Zhang 1992), 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. If only one iteration of EM is run in step (2), the resulting algorithm is the one proposed by Zhang (1992).

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 Besag (1986) 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 Geman and Geman (1984) 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 Besag 1986) 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, Dang (1998) 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 Besag 1986 and Wu and Doerschuk 1995), 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 (Geman and Geman 1984). 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 Besag (1986), 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, Qian and Titterington (1991) 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

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

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

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

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

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

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 (Celeux and Govaert 1992) or the Stochastic EM (SEM) algorithm (Celeux and Diebolt 1985). 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 Besag (1986). It is recovered exactly if, in the E step, $P_{\tilde{\mathbf{Z}}^{(q)}}(\mathbf{z} \mid y, \Psi^{(q-1)})$ 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. 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, as for the PPL-EM algorithms of Qian and Titterington (1991). 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,

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

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

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

See Appendix 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 estimate 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 paper. For comparison we also provide the results when applying the EM algorithm for independent mixture model, and when using ICM. The EM algorithm for mixtures 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.

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.

We investigated three examples. We first compared the algorithms in terms of parameter estimation, on simulated data (Section 6.1). In the two other cases, we studied the performance, in terms of the quality of the segmentations, when the true image is known (Section 6.2) and when the classification has to be consistent with some a priori knowledge (climatology example of Section 6.3).

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

6.1 Hidden 2-color Potts models

We first tested our 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 Swendsen and Wang 1987) 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. The algorithms were initialized using the same segmentation computed by simple thresholding. We divided the pixel values range, in the degraded image, into regular intervals and assigned each of them to a component. 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. 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 approach (MAP rule) for identifying the hidden labels $\mathbf{z}$ and consequently tends to produce biased Gaussian parameters estimates (see for instance Titterington 1984).

6.2 Simulated degradations of K-color images

We created a 128×128 image by adding some Gaussian noise to the 4-color image (a) of Figure 1, leading to Figure 1 (b). For this data, the hidden field model is unknown but the simulated noise corresponds to $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, 4\}$ with $\mu_k = k$ and $\sigma_k = 0.5$ for $k = 1, \ldots, 4$. We considered a model with second order neighborhood, (*i.e.* the eight closer neighbors for each pixel), which is in general, more realistic. In order to test the procedures ability to recover that the variances were equal, we tried to fit a model with class dependent variances. Our procedures, as well as EM and ICM, were applied to obtain classifications in Figure 1 (d) to (f). 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. All the algorithms were initialized by the same segmentation obtained by thresholding as in Section 6.1. The common variance is always recovered, but the algorithms using spatial information clearly outperform EM for independent mixture model in terms of restoration. Our procedures provide similar estimations of parameter β, which was not the case for the previous synthetic images we analysed. Moreover, they produce notably lower error rates than ICM.

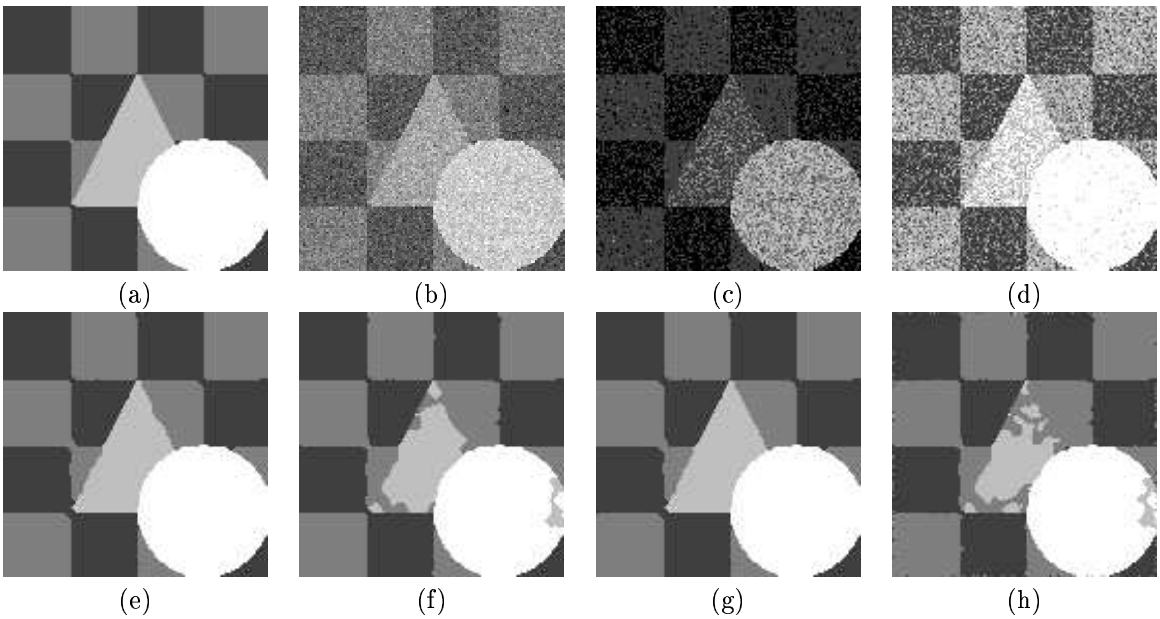

Figure 1: A 4-color image degraded with noise. (a) original image, (b) simulated image obtained by superposition of Gaussian noise with standard deviation $\sigma = 0.5$, (c) initial segmentation using simple thresholding, (d) EM segmentation, (e) mean field segmentation, (f) mode field segmentation, (g) simulated field segmentation, (h) ICM segmentation .

	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	-	1.0	2.0	3.0	4.0	0.5	0.6	0.5	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: Parameter estimates and error rates for the degraded 4-color image.

	β
Mean Field	1.9
Mode Field	1.6
Simulated Field	1.6
ICM	1.7

Table 3: Estimates of β for the precipitation climatology data.

6.3 Precipitation Climatology

Input data are data for a global precipitation climatology that has been produced at the Joint Institute for the Study of the Atmosphere and Ocean; they are available on the Web at `http://tao.atmos.washington.edu/legates_msu`. The spatial resolution of this climatology is 2.5 degrees in each of latitude and longitude, which leads to a set of twelve 144×72 maps representing stations or points (pixels) at which monthly average precipitation (in mm) has been recorded or extrapolated, for each individual calendar month. Figure 2 shows such a map for the month of January.

Possible goals of classifications of these data into a small number of components include building climatic regionalizations to show climatic variability, or defining local forecast zones consisting of groups of stations, each of which would be considered a single locale for forecasting purposes, thus reducing the total number of stations. More background can be found in Fovell (1997) where similar data have been considered for the purpose of delineating climate zones of the conterminous United States. In that study, standard clustering techniques that do not take spatial location and dependence into account were used. They have the disadvantage of producing small separate entities which are not climatologically meaningful. Using our procedures to classify these data has the advantage of producing more spatially cohesive regionalizations.

We computed six-component classifications following a suggestion in Fovell (1997). As these data are far from normally distributed, a non-linear transform was first applied: the power 0.25 of each record was taken. A fast first segmentation of the 144×72 12-band image was then obtained using the technique described by Posse (2000) (Figure 3 (a)). For the Potts model, a second order neighborhood has been used. Classification results are shown in Figure 3 and the corresponding estimates of β are given in Table 3.

Note that, depending on the goal of the analysis, the data may be pre-processed differently. If the goal is to construct local forecast zones, it may be judged prudent to standardize the records to eliminate level (mean) and seasonality (variance) distinctions. In the present application, we were more concerned with the construction of climatic zones for which the level and seasonality components are useful information. Time series can be computed by averaging the records of the members in each class, for each segmentation in Figure 3. Figure 4 shows such series for segmentations (c) to (f). We did not include the EM segmentation, Figure 3 (b), because the climatic zones do not seem relevant. Visually, segmentations (c), (d), (f) (mean, mode field algorithms and ICM segmentations) are similar. In the classifications, class 1 is represented in white while class 6 is in black. Class 1 and 2 correspond to dry regions with different seasonalities. Class 6 corresponds to wet areas with light seasonal variations. Class 3 includes regions with high variability and dry-summer climates, while class 5 includes continental interiors which generally exhibit a wet-summer/dry-winter cycle. Class 4 is characterized by moderate rainfall all year round, with a slight peak in late summer and fall. This includes much of the industrialized world: most of Europe, eastern North America, eastern Australia and, arguably, Japan, as well as large parts of southern Asia and both major oceans. This seems to be a rather heterogeneous class, suggesting that more than six classes may be needed.

The segmentation obtained with the simulated field algorithm (Figure 3 (e)), shows the same main features but clusters belonging to classe 6 are missing in some of the oceans and class 1 is much less represented in North-Africa to the favor of class 2. However the averaged time series are not significatively different suggesting that these pixels may be difficult to classify.

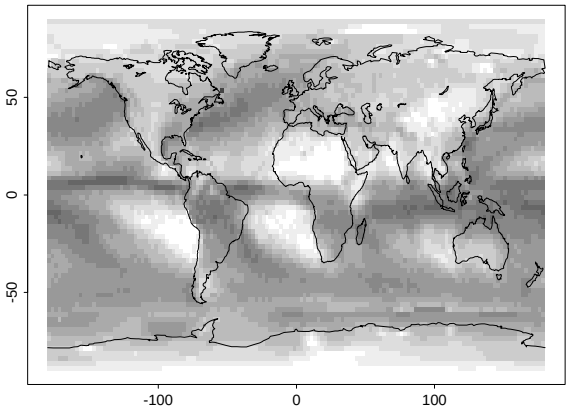

Figure 2: Monthly Average Precipitation (mm) for the month of January. The spatial resolution is 2.5 degrees in each of latitude and longitude and results in a 144×72 image.

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. 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 and is very closed to a procedure proposed by Zhang (1992). The *mode field* algorithm is closer to schemes like the ICM algorithm of Besag (1986), while the *simulated field* algorithm is related to the Gibbs sampler of Geman and Geman (1984). Comparing with the *non spatial* EM algorithm for independent mixture models and with the ICM algorithm, we observed, most of the time, significant improvement in using our algorithms, dealing with spatial dependence through approximations. In particular, the simulated field algorithm showed good performance. Its relationship to the Gibbs sampler and Monte-Carlo EM-like algorithms (Wei and Tanner 1990 and Comer and Delp 2000) suggests similar behavior in a much smaller computing time. In addition, we observed no notable time differences for running one iteration of the mean, mode, simulated field algorithms and the ICM algorithm.

Possible extension of our work is the use of other models for the unobserved image. The basic Potts models (no external field), although performing surprisingly 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 Descombes, Mangin, Pechersky, and Sigelle (1995) and Tjelmeland and Besag (1998). 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.

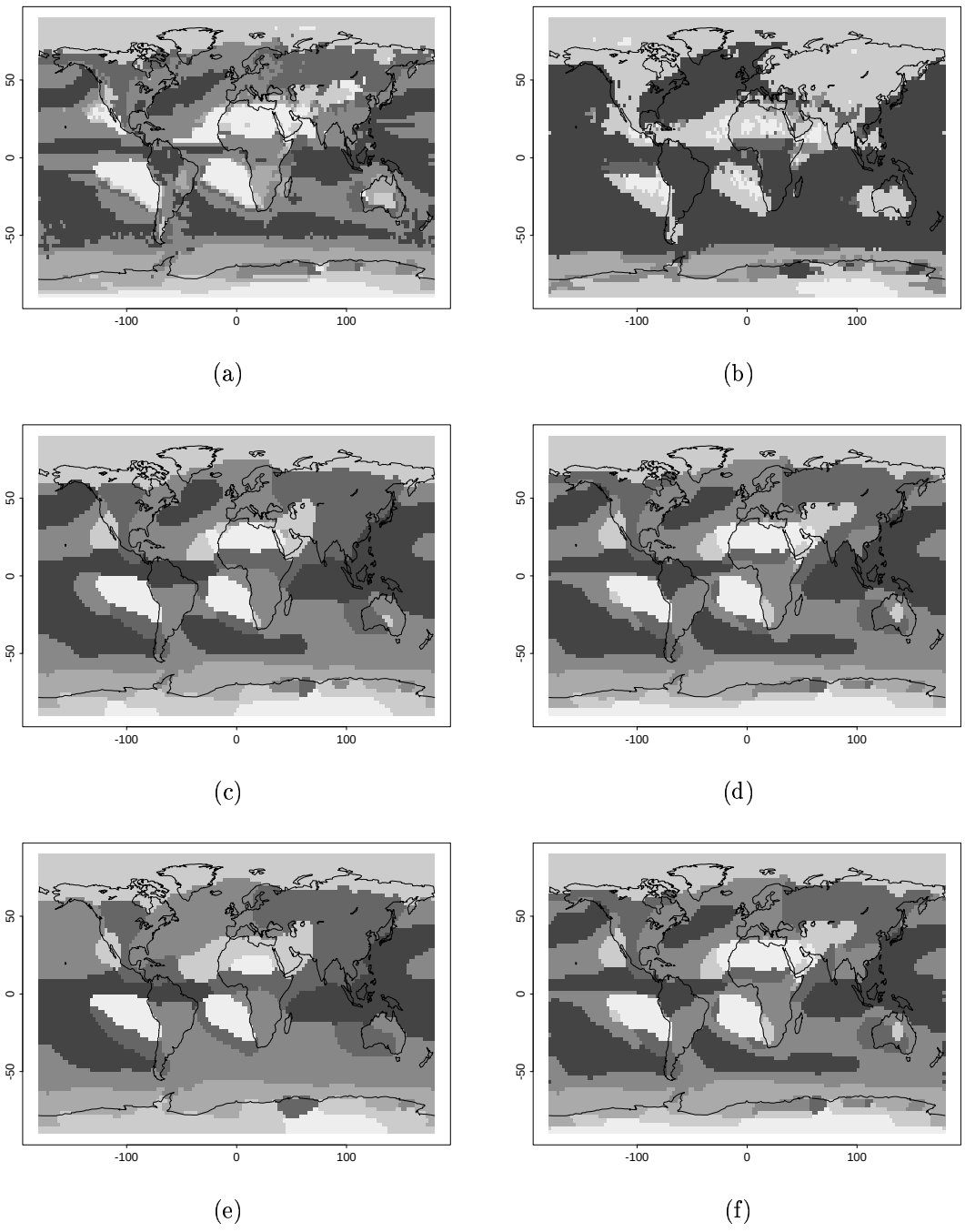

Figure 3: Segmentation of the Legates/MSU Precipitation 12-band Image: (a) initial segmentation, (b) EM segmentation, (c) mean field segmentation, (d) mode field segmentation, (e) simulated field segmentation, (f) ICM segmentation.

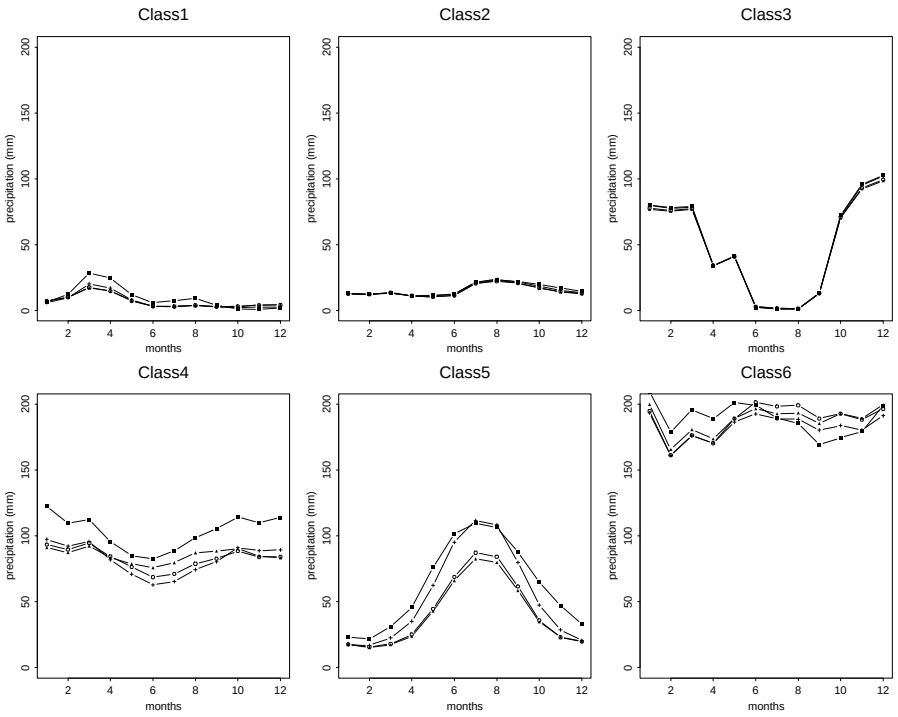

Figure 4: Time series computed by averaging the data for the members in each class of Figures 3 (c) to (f). Plotting symbols are crosses for the mean field segmentation, circles for the mode field segmentation, squares for the simulated field segmentation and triangles for the ICM segmentation.

References

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

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24*, 179–195.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, series B 48*, 259–302.

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

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

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

Chandler, D. (1987). *Introduction to Modern Statistical Mechanics.* Oxford University Press.

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

Dang, V. M. (1998). *Classification de Données Spatiales: Modèles Probabilistes et Critères de Partionnement.* Ph. D. thesis, Université de Technologie de Compiègne, France.

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

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

Fovell, R. G. (1997). Consensus clustering of U.S. temperature and precipitation data. *Journal of Climate 10*, 1405–1427.

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

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

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

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

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models.* Wiley.

Parisi, G. (1988). *Statistical Field Theory.* Addison-Wesley.

Posse, C. (2000). Hierarchical Model-Based Clustering for Large Data Sets. *To appear in Journal of Computational and Graphical Statistics.*

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

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

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

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

Wei, G. and M. A. Tanner (1990). 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.

Wu, C.-H. and P. Doerschuk (1995). 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.

Yuille, A. L. (1990). Generalized deformable models, statistical physics and matching problems. *Neural Computation 2*, 1–24.

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

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

Appendix: Mean field approximation for Potts models

To illustrate the mean field approximation principle for the Markov random fields (1) of Section 2.1, we write into more detail the equations for models with only pairwise interactions, that is with a neighborhood system such that cliques contain only one or two sites. The energy function (4) can be written

$$H(\mathbf{z}) = \sum_i \left[V_{\{i\}}(z_i) + 1/2 \sum_{j \in N(i)} V_{\{i,j\}}(z_i, z_j) \right] ,$$

assuming that $V_{\{i,j\}}(z_i, z_j) = V_{\{i,j\}}(z_j, z_i)$. The factor $1/2$ is to avoid counting twice the same sites. Assuming further that each z_i takes one of K states, each of them can be represented by a binary vector of length K with one component being 1, all others being 0. The K possible vectors are denoted by $e_1, \ldots . e_K$. A general expression for the energy function is

$$H(\mathbf{z}) = \sum_i \left\{ z_i^t V_{\{i\}} + 1/2 \sum_{j \in N(i)} z_i^t V_{\{i,j\}} z_j \right\} ,$$

where $V_{\{i\}}$ is a K-dimensional vector whose kth component is $V_{\{i\}}(e_k)$ and $V_{\{i,j\}}$ is a $K \times K$ matrix whose $(k,l)th$ component is $V_{\{i,j\}}(e_k, e_l)$. For this pairwise interaction model, the mean field consistency condition (15) is

$$\forall i \in S \ , \overline{z}_i = \frac{\sum_{z_i} z_i \exp\{-z_i^t (V_{\{i\}} + \sum_{j \in N(i)} V_{\{i,j\}} \overline{z}_j)\}}{\sum_{z_i} \exp\{-z_i^t (V_{\{i\}} + \sum_{j \in N(i)} V_{\{i,j\}} \overline{z}_j)\}} .$$

It is often convenient to rather use formulation (24) below, in terms of probability distributions. For binary vectors of length K, we have

$$\forall i \in S, \quad \bar{z}_i = I\!E_i^{mf}[Z_i] = \begin{cases} P_i^{mf}(e_1) \\ \vdots \\ P_i^{mf}(e_K) \end{cases} ,$$

and $\forall i \in S, s = 1, \ldots, K$,

$$P_i^{mf}(e_s) = \frac{\exp\{-[V_{\{i\}}(e_s) + \sum_{j \in N(i)} \sum_{l=1}^{K} V_{\{i,j\}}(e_s, e_l)\ P_j^{mf}(e_l)]\}}{\sum_{k=1}^{K} \exp\{-[V_{\{i\}}(e_k) + \sum_{j \in N(i)} \sum_{l=1}^{K} V_{\{i,j\}}(e_k, e_l)\ P_j^{mf}(e_l)]\}} . \tag{24}$$

Using the notation in Section 3, the mean field approximation of $P_G(\mathbf{z})$ is

$$\begin{aligned} P^{mf}(\mathbf{z}) &= {W^{mf}}^{-1} \exp(-H^{mf}(\mathbf{z})) \\ &= \prod_{i \in S} {W_i^{mfloc}}^{-1} \exp(-H_i^{mfloc}(z_i)) , \end{aligned}$$

where, in the pairwise interaction case, $H_i^{mfloc}(z_i)$ defined as in (13) becomes

$$H_i^{mfloc}(z_i) = z_i^t \{V_{\{i\}} + \sum_{j \in N(i)} V_{\{i,j\}} \bar{z}_j\} ,$$

and the normalizing constant is

$$W_i^{mfloc} = \sum_{z_i} z_i \exp(-z_i^t\{V_{\{i\}} + \sum_{j\in N(i)} V_{\{i,j\}}\bar{z}_j\}) .$$

In the next two examples, we specify further by considering some of the most commonly used models in image segmentation.

Example 1 (K-color Potts model with no external field) *The K-color Potts model corresponds to $V_{\{i,j\}} = -\beta Id$ for all i and j and $V_{\{i\}} = 0$ for all i (no external field), where β is a real, non-negative parameter. Notation $i \sim j$ means that sites i and j are neighbors. The model, also referred to as an homogeneous isotropic Potts model, is given by*

$$P_G(\mathbf{z} \mid \beta) = W(\beta)^{-1} \exp(\beta \sum_{i\sim j} z_i^t z_j) , \tag{25}$$

and the consistency condition becomes

$$\forall i \in S , \quad \bar{z}_i = \frac{\sum_{z_i} z_i \exp\{\beta z_i^t \sum_{j\in N(i)} \bar{z}_j\}}{\sum_{z_i} \exp\{\beta z_i^t \sum_{j\in N(i)} \bar{z}_j\}} . \tag{26}$$

A solution of these equations (26) is the uniform configuration $\bar{z}_i = (K^{-1}, \ldots, K^{-1})^t$ for all i in S. Denoting by N the number of neighbors in the model, it comes that this fixed point is stable if $\beta < \frac{K}{N}$ and is not stable if $\beta > \frac{K}{N}$. We think, but were able to prove it only for $K \leq 4$, that this fixed point is unique if $\beta < \frac{K}{N}$. For $K = 2$, the Potts model is equivalent to the Ising model for which it is known that when $\beta < \frac{2}{N}$ there is a unique fixed point and when $\beta > \frac{2}{N}$ there are two symmetric ones (see for instance Parisi 1988).

Note that regarding the Potts model with no external field, the mean field approximation may not be of great interest due to the existence of a uniform solution independently of parameter β. It appears that the resulting approximation of the marginal distribution can lead to some inconveniences when estimating β using the framework we propose in Section 4.1. The updating of β in step (2) of the general procedure is given by the maximization of (21). Suppose that $\tilde{\mathbf{z}}$ is such that $\tilde{z}_i$ does not depend on i in S. This assumption is natural since there is no external field in the model. Let $\mathbf{m} = (m_1, m_2, \ldots, m_K)$ be this common value, and $P_{\mathbf{m}}$ the corresponding approximation, (21) becomes

$$\beta^{(q)} = \arg\max_{\beta} N\beta \sum_{k=1}^{K} m_k \sum_{i\in S} P_{\mathbf{m}}(e_k \mid y, \Psi^{(q-1)}) - |S| \log(\sum_{k=1}^{K} e^{N\beta m_k}). \tag{27}$$

If in addition, $\mathbf{m}$ has all its components equal (necessarily to K^{-1}), then the function to maximize in (27) reduces to $-S \log K$, which does not depend on β, so that the parameter cannot be updated. As mentioned above, such a $\tilde{\mathbf{z}}$ is always solution of equation (26). Then, algorithms in our framework using a mean field approximation of the marginal distribution to compute $\tilde{\mathbf{z}}$ would not lead to sensible estimation of β. Similar remarks disuading of using the mean field approximation for an homogeneous isotropic Potts model can be found in Archer and Titterington (2000).

In practice, this is not a problem for image analysis since the observed image can be seen as an external field. So that the Potts model involved is the one given in the next example with $V_{\{i\}} = -\log f_i(y_i \mid z_i, \theta)$ for all i in S. For such a model, there is usually no uniform solution of the corresponding fixed point equation.

Example 2 (K-color Potts model with site dependent external field) *The K-color Potts model considered here, corresponds to $V_{\{i,j\}} = -\beta Id$ for all i and j (Potts model (25)) and $V_{\{i\}}$ different from zero (site dependent external field). For this model,*

$$P_G(\mathbf{z} \mid \beta) = W(\beta)^{-1} \exp\{-\sum_{i\in S} z_i^t V_{\{i\}} + \beta(\sum_{i\sim j} z_i^t z_j)\} .$$

When $K = 2$, *we can write* $V_{\{i\}} = (v_i(1), v_i(2))^t$ *and* $p_i = P_i^{mf}(e_1) = 1 - P_i^{mf}(e_2)$. *Then equation (24) becomes*

$$\forall i \in S, \quad p_i = [1 + \exp\{v_i(1) - v_i(2) + \beta(|N(i)| - 2 \sum_{j \in N(i)} p_j)\}]^{-1} .$$

The approximation of $I\!E_G(Z_i)$ *is*

$$\bar{z}_i = \begin{cases} p_i \\ 1 - p_i \end{cases} .$$

For i *in* S, *the mean field approximation of* $P_G(z_i \mid \beta)$ *is* $p_i^{z_i^t e_1}(1 - p_i)^{z_i^t e_2}$ *and the mean field approximation of* $P_G(\mathbf{z} \mid \beta)$ *is then*

$$\prod_{i \in S} p_i^{z_i^t e_1}(1 - p_i)^{z_i^t e_2} .$$

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399