Hidden Markov Random Field Model Selection Criteria Based on Mean Field-Like Approximations

Florence Forbes and Nathalie Peyrard

Abstract—Hidden Markov random fields appear naturally in problems such as image segmentation, where an unknown class assignment has to be estimated from the observations at each pixel. Choosing the probabilistic model that best accounts for the observations is an important first step for the quality of the subsequent estimation and analysis. A commonly used selection criterion is the Bayesian Information Criterion (BIC) of Schwarz (1978), but for hidden Markov random fields, its exact computation is not tractable due to the dependence structure induced by the Markov model. We propose approximations of BIC based on the mean field principle of statistical physics. The mean field theory provides approximations of Markov random fields by systems of independent variables leading to tractable computations. Using this principle, we first derive a class of criteria by approximating the Markov distribution in the usual BIC expression as a penalized likelihood. We then rewrite BIC in terms of normalizing constants, also called partition functions, instead of Markov distributions. It enables us to use finer mean field approximations and to derive other criteria using optimal lower bounds for the normalizing constants. To illustrate the performance of our partition function-based approximation of BIC as a model selection criterion, we focus on the preliminary issue of choosing the number of classes before the segmentation task. Experiments on simulated and real data point out our criterion as promising: It takes spatial information into account through the Markov model and improves the results obtained with BIC for independent mixture models.

Index Terms—Image segmentation, hidden Markov random fields, model selection, Bayesian Information Criterion, mean field approximation, partition function.

1 Introduction

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 at each pixel of an image while the hidden data could consist of an unknown class assignment to be estimated from the observations at each pixel. This case is usually referred to as image segmentation. In the context of statistical image segmentation, choosing the probabilistic model that best accounts for the observations is an important first step for the quality of the subsequent estimation and analysis. In most cases, the choice is done subjectively using expert knowledge or ad hoc procedures and there is a striking lack of systematic data-based approaches. We recast this choice as a problem of probabilistic model comparison and use the standard approach of Bayes factors. Evaluating the Bayes factor of one model against another involves calculating the ratio of the integrated likelihoods for each model, i.e., the likelihoods of the observations integrated over the respective model parameters. For a lot of models of interest, these integrated likelihoods are high-dimensional and intractable integrals so that most available software is generally inefficient for their evaluation. Various approximations have been proposed. In particular, the Bayesian Information Criterion (BIC) approximation of [1] is based on the Laplace method for integrals. It leads to an equation giving the log-integrated likelihood as the maximized log-likelihood minus a correction or penalization term and an $O(1)$ error term, as the sample size tends to infinity. BIC can be compared to other selection criteria. One of them is the Akaike Information Criterion (AIC) of [2], which differs from BIC in the correction term, but has been shown to overestimate the number of parameters in practice. The criterion proposed in [3] is based on stochastic complexity and is similar to BIC, and methods using cross validation [4] seem promising, but their tractability in our context is not straightforward due to the dependence structure in the data. Many other approaches can be found in the literature on model selection (see for instance the list of references in [5]).

BIC has become quite popular due to its simplicity and its good results in cases where p-values and the standard model selection procedures based on them were unsatisfactory. P-values (see [6]), or observed significant levels, are indicators of the strength of the evidence of one model against another in hypothesis testing, but can be highly misleading when used for model selection (see [7]). In BIC, the $O(1)$ error does suggest the approximation to be somewhat crude. However, empirical experience has found the approximation to be more accurate in practice than the $O(1)$ error term would suggest. As regards model selection, Kass and Raftery [5] observe that the criterion does not seem to be grossly misleading in a qualitative sense as long

- *F. Forbes is with Projet IS2, INRIA Rhône-Alpes, ZIRST, 655 av. de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France. E-mail: Florence.Forbes@inrialpes.fr.*
- *N. Peyrard is with Projet VISTA, IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France. E-mail: npeyrard@irisa.fr.*

Manuscript received 6 Feb. 2002; revised 26 Aug. 2002; accepted 17 Dec. 2002.
Recommended for acceptance by P. Meer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 115839.

as the number of degrees of freedom involved in the comparison is relatively small relative to sample size. In this paper, we consider Markov model-based image segmentation and focus on the use of BIC for the underlying issue of choosing a model from a collection of hidden Markov random fields. In this case, we have no specific results on the quality of BIC as an approximation of the integrated likelihood and this choice as a selection criterion is arguable. However, the question of the criterion ability to asymptotically choose the correct model can be addressed independently of the integrated likelihood approximation issue. As an illustration, the author in [8] have proven recently that for the more specialized but related case of hidden Markov chains, under reasonable conditions, the *maximum penalized marginal likelihood* estimator of the number of hidden states in the chain is consistent. This estimator is defined for a class of penalization terms that includes the BIC correction term and involves an approximation of the maximized log-likelihood which is not necessarily good, namely the maximized log-marginal likelihood. In particular, this criterion is consistent even if there is no guarantee that it provides a good approximation of the integrated likelihood. The choice of BIC for hidden Markov model selection appears then reasonable and we will show that criteria with good experimental behavior can be derived from it.

The difficulty in the context of hidden Markov random fields lies in that the maximized log-likelihood part in BIC involves Markov distributions whose exact computation requires an exponential amount of time. As regards observed Markov random fields selection, Ji and Seymour [9] propose a consistent procedure based on penalized Besag pseudolikelihood [10], [11] study a Markov Chain Monte Carlo (MCMC) approximation of BIC. When the fields are hidden, little has been done to address the selection problem. Two approximations of BIC are proposed in [12]: For the Pseudo-Likelihood Information Criterion (PLIC) the required maximized distribution is approximated by the Qian and Titterington pseudolikelihood [13], while a simpler approximation, the Marginal Mixture Information Criterion (MMIC) is based on the marginal distribution of pixel values. In practice, good results are reported for PLIC in [12], whereas MMIC is less satisfactory. In this paper, we propose approximations of BIC based on the mean field principle. Mean field theory of statistical physics [14] is an approach providing an approximation of a Markov random field by a system of independent variables and leading to tractable computations. We use a generalization of the mean field principle presented in a previous work [15] and derive a class of criteria that includes PLIC as a particular case and as a result gives some new insight on its nature. We also show that the straightforward use of the mean field approximation can be improved by rewritting BIC in terms of normalizing constants, also called partition functions, instead of Markov distributions and then using optimal mean field lower bounds, usually referred to as Gibbs-Bogoliubov-Feynman bounds, for the normalizing constants. We derive this way another tractable criterion denoted by BIC^{GBF}. Questions of interest relevant to model selection include choosing the Markov field neighborhood or more generally its energy function and choosing the number of classes in which to segment the data. They can all be addressed straightforwardly in our framework, but we focus on the latter because of its practical importance. Experiments on simulated and real data point out BIC^{GBF} as a promising criterion. It is easy to compute and shows good and stable performance. It takes spatial information into account through the Markov model and improves the results obtained with BIC for independent mixture models. In particular, it seems to avoid the overestimation of the number of classes observed in [16].

The complete parametric models for the observed and unobserved variables are specified in Section 2 and the basics for BIC are recalled in Section 3. The mean field approximation principle is briefly presented in Section 4 and in Section 5, we show how we propose to use it to compute approximations of BIC and derive new computationally tractable criteria for hidden Markov model selection. Experiments are reported in Section 6 and a discussion section ends the paper.

2 Hidden Markov Models

Let S be a finite set of sites with a neighborhood system defined on it. Let $|S| = n$ 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 the sites are all neighbors. Let $V = \{e_1, \ldots, e_K\}$ 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, (1) means that interactions between site i and the other sites actually reduce to interactions with its neighbors. Equation (2) is important for the Hammersley-Clifford Theorem [17] (or [18] for a published reference) 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 sum is over the set of cliques and the V_cs are the clique potentials which may depend on parameters not specified in the notation $W = \sum_{\mathbf{z}} \exp(-H(\mathbf{z}))$ is the normalizing constant also called the partition function. We will write $\sum_{\mathbf{z}}$ (respectively, $\sum_{\mathbf{z}_A}$) a sum over all possible values of $\mathbf{z}$ (respectively, $\mathbf{z}_A$). 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 distribution (3) is the pseudolikelihood introduced by [10] and defined as

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

Each term in the product is easy to compute. For a given value z_i of variable Z_i,

$$P_G(z_i \mid \mathbf{z}_{N(i)}) = \frac{\exp(-\sum_{c\,,\, i \in c} V_c(\mathbf{z}_c))}{\sum_{z_i'} \exp(-\sum_{c\,,\, i \in c} V_c(\mathbf{z}'_c))}, \tag{6}$$

where the sums in the exponentials are only over cliques c that contain site i and where the outer sum in the denominator is over all possible values z_i' for Z_i. For c containing i and a given z_i', $\mathbf{z}'_c$ denotes values $\{z_i', z_j, j \in c, j \neq i\}$ for sites in c. Equation (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. It has been used by [19] in the model selection context (see Section 5). In Section 5, we will use other approximations based on systems of independent variables. Their factorization properties simplify computations as does approximation (5) and they correspond to valid probability models.

Image segmentation involves observed variables and unobserved variables to be recovered. The unobserved variables are 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_is are given),

$$\begin{aligned} f(\mathbf{y} \mid \mathbf{z},\ \theta) &= \prod_{i \in S} f_i(y_i \mid z_i,\ \theta) \\ &= \exp\left\{\sum_{i \in S} \log f_i(y_i \mid z_i,\ \theta)\right\}, \end{aligned} \tag{7}$$

assuming that all the $f_i(y_i \mid z_i,\ \theta)$ are positive. This makes the model similar to an independent mixture model [20]. 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

$$\begin{aligned} P_G(\mathbf{y}, \mathbf{z} \mid \theta, \beta) &= f(\mathbf{y} \mid \mathbf{z}, \theta)\ P_G(\mathbf{z} \mid \beta) \\ &= W(\beta)^{-1} \prod_{i \in S} f_i(y_i \mid z_i,\ \theta)\ \prod_c \exp\{-V_c(\mathbf{z}_c \mid \beta)\} \\ &= W(\beta)^{-1} \exp\left\{-H(\mathbf{z} \mid \beta) + \sum_{i \in S} \log f_i(y_i \mid z_i,\ \theta)\right\}. \end{aligned} \tag{8}$$

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

$$H(\mathbf{z} \mid \mathbf{y}, \theta, \beta) = H(\mathbf{z} \mid \beta) - \sum_{i \in S} \log f_i(y_i \mid z_i,\ \theta). \tag{9}$$

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 and denote by $\Psi = (\theta, \beta)$ the parameter vector.

3 Bayesian Information Criterion

In a Bayesian framework, a way of selecting a model among R models $M_1, M_2, \ldots, M_R$ consists of choosing the model with highest posterior probability. By Bayes theorem, the posterior probability of M_r ($r \in \{1, \ldots, R\}$) given the observations y is

$$P(M_r|\mathbf{y}) = \frac{P_G(\mathbf{y}|M_r)P(M_r)}{\sum_{k=1}^{R} P_G(\mathbf{y}|M_k)P(M_k)},$$

where $P_G(\mathbf{y}|M_r)$ is the integrated or marginal likelihood of model M_r and $P(M_r)$ is its prior probability. Assuming that all models have equal prior probabilities, choosing the model with the highest posterior probability is equivalent to select the model with the largest integrated likelihood,

$$P_G(\mathbf{y}|M_r) = \int P_G(\mathbf{y}|\Psi_r, M_r)P(\Psi_r|M_r)d\Psi_r, \tag{10}$$

where Ψ_r varies in the model M_r parameter space and $P(\Psi_r|M_r)$ is the prior distribution on Ψ_r for the same model. Computing (10) is not usually tractable. A simple and often reliable way to approximate the integrated likelihood is provided by the Bayesian Information Criterion (BIC) of [1] (see, for instance, [5]),

$$2 \log P_G(\mathbf{y}|M_r) \approx \mathrm{BIC}(M_r) = 2 \log P_G(\mathbf{y} \mid \Psi_r^{ml}) - d_r \log(n), \tag{11}$$

where Ψ_r^{ml} is the maximum-likelihood estimate of Ψ_r,

$$\Psi_r^{ml} = \arg\max_{\Psi_r} P_G(\mathbf{y} \mid \Psi_r, M_r),$$

d_r is the number of free parameters in model M_r and n is the number of observations ($n = |S|$ the number of sites). It has been widely used in the context of selecting the number of components in independent mixture models [21], [22]. In this context, BIC limitations have been pointed out. In particular, it has been observed that in practice the criterion can tend to overestimate the right number of components when the true model is not in $\{M_1, \ldots . M_R\}$ [16].

For hidden Markov models, the difficulty comes from that Ψ_r^{ml} and $P_G(\mathbf{y} \mid \Psi_r^{ml})$ are not available. For computing BIC, methods using simulations have been investigated in [23], while [19] proposed using the pseudolikelihood (5) as an approximation to the intractable Markov distribution. In this paper, we suggest using the mean field approximation principle to derive a class of other tractable criteria. As for the pseudolikelihood approximation, it consists of replacing the original Markov distribution by a product easier to deal with. We recall the mean field principle in the next section and describe applications in the model selection context in Section 5.

4 Mean Field Theory

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 physics [14] where it has been used to study phase transition phenomena. More recently, it has been used in computer vision applications [24], [25], [26], graphical models [27] and references therein, and other areas [28]. This principle provides an approximation of the distribution P_G of a Markov random field. The idea when considering a particular site i is to neglect the fluctuations of the sites interacting with i, by fixing them to their mean values. The resulting system behaves as one composed of independent variables, with factorized distribution, for which computation gets tractable. Let $\mathbb{E}_P$ denote the expectation under distribution P. A proper presentation (see, for instance, [28]) and a rational for using the mean field approximation arise from the minimization of the Kullback-Leibler divergence, $KL[P, P_G] = \mathbb{E}_P\left[\log\left(\frac{P(Z)}{P_G(Z)}\right)\right]$, between a given distribution P and the Gibbs distribution P_G, over the set of probability distributions $P = \prod_{i\in S} P_i$ that factorize. The Kullback-Leibler divergence is a measure of dissimilarity between two distributions. It is always positive and is zero only when the distributions are equal. The mean field approximation of P_G, denoted by P^{mf} in what follows, is then defined as the distribution that factorizes, which is the closest to P_G in term of the Kullback-Leibler divergence. In practice, the mean field approximation $P^{mf} = \prod_{i\in S} P_i^{mf}$ is obtained by solving a fixed point equation determining its marginal distributions P_i^{mf} for all i in S. Indeed, a distribution P that factorizes into $\prod_{i\in S} P_i$ is completely defined by its marginal distributions P_i for all i in S. In our settings, these marginal distributions are defined over a finite set V of indicator vectors of length K denoted by $\{e_1, \ldots, e_K\}$ (see Section 2), each of them representing a possible value for variable Z_i. Therefore, P_i is completely defined by its values on V, i.e., by $P_i(e_1), \ldots, P_i(e_K)$ or equivalently by its expectation $\mathbb{E}_{P_i}[Z_i] = (P_i(e_1), \ldots, P_i(e_K))^t$ denoted by $\bar{z}_i$ in what follows. Then, finding the factorized distribution P that minimizes $KL[P, P_G]$ consists of finding, for all i in S, the optimal $\bar{z}_i$s. Computing the gradient of the Kullback-Leibler divergence with regards to the $\bar{z}_i$s and setting it to zero (see [14] for details), leads to the following fixed point equation involving $\bar{\mathbf{z}} = \{\bar{z}_j, j \in S\}$ and P_G,

$$\bar{\mathbf{z}} = g(\bar{\mathbf{z}}) \;=\; \begin{cases} g_1(\{\bar{z}_j, j \in N(1)\}) \\ \vdots \\ g_n(\{\bar{z}_j, j \in N(n)\}), \end{cases} \tag{12}$$

where for all i in S, $g_i(\{\bar{z}_j, j \in N(i)\}) = \sum z_i P_G\big(z_i \mid \bar{\mathbf{z}}_{N(i)}\big)$ with the sites in $|S|$ numbered from 1 to n.[2]The mean field approximation consists of solving this fixed point equation and taking the solution denoted by $\mathbf{z}^{mf} = \{z_i^{mf}, i \in S\}$ as an estimate of the exact mean field $\mathbb{E}_{P_G}[\mathbf{Z}]$. In the right-hand side of (12), $g_i(\{z_j^{mf}, j \in N(i)\})$ is the expectation of Z_i under the conditional distribution $P_G(\,.\mid \mathbf{z}_{N(i)}^{mf})$, that is intuitively, under the original Gibbs distribution P_G where the neighbors (sites in $N(i)$) are fixed to $\mathbf{z}_{N(i)}^{mf}$. We can recover this way the interpretation of mean field approximation as a way to deal with the interactions in the original Gibbs measure P_G by setting the neighbors to their mean values. Also, (12) states that the mean field computed based on the approximation (right-hand side) is equal to the mean field used to define this approximation (left-hand side). It is often referred to as a self-consistency condition. The mean field approximation $P^{mf}(\mathbf{z})$ of the Gibbs distribution is then defined by

$$P^{mf}(\mathbf{z}) = \prod_{i\in S} P_i^{mf}(z_i), \tag{13}$$

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

It follows straightforwardly an expression of P^{mf} as a Gibbs distribution,

$$P^{mf}(\mathbf{z}) = \frac{1}{W^{mf}} \exp(-H^{mf}(\mathbf{z})), \tag{14}$$

where H^{mf} and W^{mf} denote, respectively, the energy function and the partition function under (13) and are easy to compute due to the factorization property. If $\mathbb{E}^{mf}$ denotes the expectation under (13), using (14), it is easy to see from the positivity of the Kullback-Leibler divergence $KL[P^{mf}, P_G]$, that the following inequality holds,

$$W \geq W^{mf} \exp(-\mathbb{E}^{mf}[H(Z) - H^{mf}(Z)]) \;. \tag{15}$$

This inequality is known as the the Gibbs-Bogoliubov-Feynman (GBF) bound [14]. Note that the same inequality is valid for any energy function other than H^{mf}. However, the mean field model (13) is optimal among models with factorization property, in the sense that it maximizes the lower bound in inequality (15) for such models. When considering the expansion around zero of $\exp(-\mathbb{E}^{mf}[H(Z) - H^{mf}(Z)])$, the right-hand side of inequality (15), denoted by W^{GBF} in what follows:

$$W^{GBF} = W^{mf} \exp(-\mathbb{E}^{mf}[H(Z) - H^{mf}(Z)]) \;, \tag{16}$$

can be seen as a first order approximation (in $\Delta H = H - H^{mf}$) of the normalization constant W (see [14]). As a first order approximation, we can expect W^{GBF} to be a closer approximation of W than W^{mf} which corresponds to the zeroth order. This is illustrated by the example in the Appendix where the three quantities W, W^{GBF}, and W^{mf} are compared for a 2-color Potts model.

Therefore, in addition to the zeroth order mean field approximation, $P_G \sim P^{mf}$, a first order approximation of the partition function W can be easily derived. In the following sections, we then propose two ways of approximating BIC. In Section 5.1, we derive BIC approximations based on approximation $P_G \sim P^{mf}$, while in Section 5.2, we show how to use the first order approximation of W.

5 Mean Field-Like Approximations of BIC

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 [15]. In Section 5.1, we apply this idea to release the computational burden when dealing with the intractable distribution $P_G(\mathbf{y} \mid \Psi)$ in BIC computation. This approach is the most straightforward, considering (11) of BIC and includes criterion PLIC introduced in [19] and recalled below. However, we will show that in practice this does not always lead to satisfying results. In Section 5.2, we show that approximating the whole distribution is actually not necessary and we derive alternative criteria approximating BIC using the first order partition function approximation (16). Experimental results confirm the superiority of this method.

As regards the notation, we consider a model M_r among R hidden Markov models ($r = 1, \ldots, R$) as defined by (3) and (7) with parameters $\Psi_r = (\theta_r, \beta_r)$.

5.1 Approximating the Gibbs Distribution

A mean field-like approximation of a Gibbs distribution can be defined as follows: Given a configuration $\tilde{\mathbf{z}}$, set the neighbors of each site i to $\tilde{\mathbf{z}}_{N(i)}$ and replace the marginal distribution $P_G(\mathbf{z} \mid \beta_r)$ by

$$P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta_r) = \prod_{i \in S} P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}, \beta_r). \tag{17}$$

It corresponds to an observed likelihood of the form

$$\begin{aligned} P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_r) &= \sum_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{z}, \theta_r) P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta_r) \\ &= \prod_{i \in S} \sum_{z_i} f_i(y_i \mid z_i, \theta_r)\ P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}, \beta_r) \\ &= \prod_{i \in S} P_G(y_i \mid \tilde{\mathbf{z}}_{N(i)}, \Psi_r). \end{aligned} \tag{18}$$

We consider $P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_r)$ as a candidate for an approximation of the intractable $P_G(\mathbf{y} \mid \Psi_r)$ involved in (11) of BIC. The flexibility of our proposition is then in the choice of the values $\tilde{\mathbf{z}}$. A natural candidate would be one that leads to a reasonable approximation of $P_G(\mathbf{y}, \mathbf{z} \mid \Psi_r)$. In our model, $P_G(\mathbf{z} \mid \beta_r)$ and $P_G(\mathbf{z} \mid \mathbf{y}, \Psi_r)$ are not available while $f(\mathbf{y} \mid \mathbf{z}, \theta_r)$ is. Knowing $f(\mathbf{y} \mid \mathbf{z}, \theta_r)$, it is enough to approximate one of the unknown quantities, either $P_G(\mathbf{z} \mid \beta_r)$ or $P_G(\mathbf{z} \mid \mathbf{y}, \Psi_r)$, to derive an approximation of the other and of the joint distribution. Therefore, our selection of $\tilde{\mathbf{z}}$ can be driven by the quality of the corresponding approximation of $P_G(\mathbf{z} \mid \beta_r)$ or $P_G(\mathbf{z} \mid \mathbf{y}, \Psi_r)$. As regards the Kullback-Leibler divergence, the approximations cannot be both optimal and satisfy the Bayes rule. 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 dissuading from using the mean field approximation on the marginal field (see [29] and [15]).

For computing BIC, it then remains the problem of computing the maximum-likelihood estimator Ψ_r^{ml}. Let $\hat{\Psi}_r$ be an approximation of Ψ_r^{ml}.

An approximation for BIC is then,

$$\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_r) = 2 \log P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \hat{\Psi}_r) - d_r \log(n). \tag{19}$$

As regards the quality of such an approximation, it is not clear whether $\tilde{\mathbf{z}}$ and $\hat{\Psi}_r$ must be chosen independently or not. As an example, the Pseudo-Likelihood Information Criterion (PLIC) of [19] is a particular case of $\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_r)$. Indeed, one possibility to get values for $\hat{\Psi}_r$ and $\tilde{\mathbf{z}}$ is to use the unsupervised Iterated Conditional Modes (ICM) algorithm of [30]. In that case, $\hat{\Psi}_r$ and $\tilde{\mathbf{z}}$ are computed using a single iterative procedure which alternates between estimating Ψ_r and estimating $\mathbf{z}$ so that the final estimates, denoted by Ψ_r^{ICM} and $\mathbf{z}_r^{ICM}$, can be deduced from one another and are not chosen independently. Then, approximation (19) becomes

$$\begin{aligned} \mathrm{BIC}^{\mathbf{z}_r^{ICM}}(\Psi_r^{ICM}) &= 2 \log(P_{\mathbf{z}^{ICM}}(\mathbf{y} \mid \Psi_r^{ICM})) - d_r \log(n) \\ &= \mathrm{PLIC}(M_r). \end{aligned} \tag{20}$$

In this paper, we propose to use for $\tilde{\mathbf{z}}$ and $\hat{\Psi}_r$ the output of the Expectation-Maximization (EM) algorithm-based procedures described in [15] and referred to as *mean field-like algorithms* in what follows. The idea underlying these algorithms is to replace the intractable Markov distribution by a simpler distribution obtained by fixing the neighbors of each pixel to constant values as in (17). Then, an iteration of a mean field-like algorithm consists of two steps: In the first step, the values $\tilde{\mathbf{z}}$ for the neighbors are updated according to the observations $\mathbf{y}$ and to the current value of the parameter. As in (17), it follows an approximation of the intractable Markov distribution. Then, the second step consists of carrying out the EM algorithm for the corresponding approximated observed likelihood (18) to obtain an updated value $\hat{\Psi}_r$ of the parameter. Mean field-like algorithms can thus be related to the EM algorithm for independent mixture models, with the significant difference that the mixture model adaptively changes at each iteration depending on the current choice of the neighbors values. Like ICM, these algorithms alternatively produce a configuration $\tilde{\mathbf{z}}$ and, using (18), an estimation $\hat{\Psi}_r$ (see Section 6 for more details). In [15], we compared three different ways of updating $\tilde{\mathbf{z}}$: the mean field approximation of the conditional mean, an approximation of the conditional mode, and a simulated realization of the conditional Gibbs distribution obtained with the Gibbs sampler of [31]. Based of this study, we focused on the last solution. It leads to an algorithm referred to in [15] as the *simulated field* algorithm, which showed good performance as regards hidden Markov random fields parameter estimation and outperformed ICM in this task in most cases.

PLIC shows promising results when used to select the number of components in tests on synthetic and real images reported in [19]. In Section 6, we report additional results for $\tilde{\mathbf{z}}$ and $\hat{\Psi}_r$ set to values provided by the ICM algorithm (PLIC). The results when $\tilde{\mathbf{z}}$ and $\hat{\Psi}_r$ are obtained via mean field-like algorithms are not reported, but can be found in [32]. They were satisfactory for real data, but surprisingly unstable as regards the number of components on simulated data (simulated Potts models as in Section 6.1).

In this first approach, the use of the simulated field algorithm for $\hat{\Psi}_r$ appears reasonable and we will keep this estimation procedure in the next section. As regards the quality of the approximation of $P_G(\mathbf{y} \mid \Psi_r)$ by $P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_r)$, it is not easy to assess but in what follows we will propose a more satisfying alternative.

5.2 Approximating the Partition Function

In this section, the idea is to use an expression for BIC that involves only partition functions so that the problem of approximating the Markov distributions can be replaced by that of approximating the partition functions. The advantage is that the partition function first order approximations presented in Section 4, can be used and results in better approximations.

Let $W(\mathbf{y}, \Psi)$ and $W(\beta)$ be the partition functions for the conditional and marginal fields, respectively,

$$W(\beta) = \sum_{\mathbf{z}} \exp(-H(\mathbf{z}|\beta))$$
$$W(\mathbf{y}, \Psi) = \sum_{\mathbf{z}} \exp(-H(\mathbf{z}|\mathbf{y}, \Psi)).$$

Using notation of Section 2, it comes from

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

that

$$P_G(\mathbf{y} \mid \Psi) = \frac{f(\mathbf{y} \mid \mathbf{z}, \theta)\ \exp(-H(\mathbf{z}|\beta))}{\exp(-H(\mathbf{z}|\mathbf{y}, \Psi))}\ \frac{W(\mathbf{y}, \Psi)}{W(\beta)},$$

which, using expression (9), simplifies into

$$P_G(\mathbf{y} \mid \Psi) = \frac{W(\mathbf{y}, \Psi)}{W(\beta)}. \tag{21}$$

Other derivations and uses of (21) can be found in other papers. For instance, in [33], (21) is used with the mean field approximation to approximate the maximum-likelihood estimator of the Markov field parameters. Using (21), the expression (11) of BIC is equivalent to the following one which uses only the partition functions $W(\mathbf{y}, \Psi)$ and $W(\beta)$,

$$\mathrm{BIC}(M_r) = 2 \log W(\mathbf{y}, \Psi_r^{ml}) - 2 \log W(\beta_r^{ml}) - d_r \log(n). \tag{22}$$

At this stage, a possible approach is to approximate the two partition functions using a Monte Carlo partition function estimation algorithm (see [34] for a unified presentation of such methods). Monte Carlo simulations, however, are very time consuming. As an alternative, we propose to use the mean field first order approximations for the partition functions. Let $H^{mf}(\mathbf{z}|\beta)$ and $H^{mf}(\mathbf{z}|\mathbf{y}, \Psi)$ denote the mean field expressions for the marginal and conditional field energies. Using the first order approximations, a new approximation of BIC is:

$$\begin{aligned}\mathrm{BIC}^{GBF}(\hat{\Psi}_r) =& 2 \log W^{mf}(\mathbf{y}, \hat{\Psi}_r) - 2\mathbb{E}^{mf}[H(\mathbf{Z}|\mathbf{y}, \hat{\Psi}_r) \\ & - H^{mf}(\mathbf{Z}|\mathbf{y}, \hat{\Psi}_r)|\mathbf{y}] \\ & - 2 \log W^{mf}(\hat{\beta}_r) + 2\mathbb{E}^{mf}[H(\mathbf{Z}|\hat{\beta}_r) \\ & - H^{mf}(\mathbf{Z}|\hat{\beta}_r)] \\ & - d_r \log(n).\end{aligned} \tag{23}$$

As before, $\hat{\Psi}_r$ must be estimated and we used mean field-like algorithms and, more specifically, the simulated field algorithm described previously. The marginal and conditional mean field approximations were then computed, using this value of the parameter, by solving the corresponding fixed point equations (12).

The expression of BIC^{GBF} (23) is more satisfactory than the approximation $\mathrm{BIC}^{\tilde{\mathbf{z}}}$ (19). A way to see the improvement is to rewrite $P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_r)$ in (18) using partition functions as in (21),

$$P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_r) = \frac{W_{\tilde{\mathbf{z}}}(\mathbf{y}, \Psi_r)}{W_{\tilde{\mathbf{z}}}(\beta_r)}.$$

Expressions for both quantities in the ratio are easily deduced from (17) and (6). The ratio (21) is thus better approximated in BIC^{GBF} than in $\mathrm{BIC}^{\tilde{\mathbf{z}}}$ since as explained in Section 4, it uses the best lower bound (15) for each partition function. Therefore, there are some theoretical and experimental reasons to believe that our BIC^{GBF} is a better approximation of the true BIC than $\mathrm{BIC}^{\tilde{\mathbf{z}}}$ and, thus, than PLIC. BIC^{GBF} is based on a better approximation of $P_G(\mathbf{y} \mid \Psi_r)$ and the procedure it uses to compute $\hat{\Psi}_r$ has shown to be as reliable if not better than ICM in practice [15]. However, note that as regards model selection, this does not necessarily ensure that the resulting criterion would lead to better results although the experiments reported in Section 6 tend to confirm this.

6 Experiments

In this section, the gain in approximating the partition functions, leading to BIC^{GBF}-like criteria, rather than the whole Markov distribution, leading to PLIC-like criteria, is investigated. We examine the performance of the two approaches as regards the problem of choosing the number of classes in the segmentation. We report experiments on three types of images. For all examples, the observed images are considered as realizations of the simple following hidden Markov model. The distribution of the hidden field is supposed to be a K-color Potts model where each z_i takes one of K states, which represent K different class assignments or colors. Recall that each of the states 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,

$$P_G(\mathbf{z} \mid \beta) = W(\beta)^{-1} \exp\left(\beta \sum_{i \sim j} z_i^t z_j\right), \tag{24}$$

where β is a real nonnegative parameter and the notation $i \sim j$ represents all pairs of sites (i, j) which are neighbors.

For the f_is we considered Gaussian distributions. If site i is in class k, f_i is the Gaussian distribution with mean μ_k and standard deviation σ_k. The parameter to be estimated is then $\Psi = \{\theta, \beta\}$ with $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$. Let M_K be the model defined above when the number of colors is K. To assess its ability to select a relevant number K, the criterion BIC^{GBF} is computed for model M_K with $K = K_{min}$ to $K = K_{max}$. The required estimations of $\hat{\Psi}_K$ for each value of K considered were obtained with the simulated field algorithm. In practice, a sequential version of this algorithm

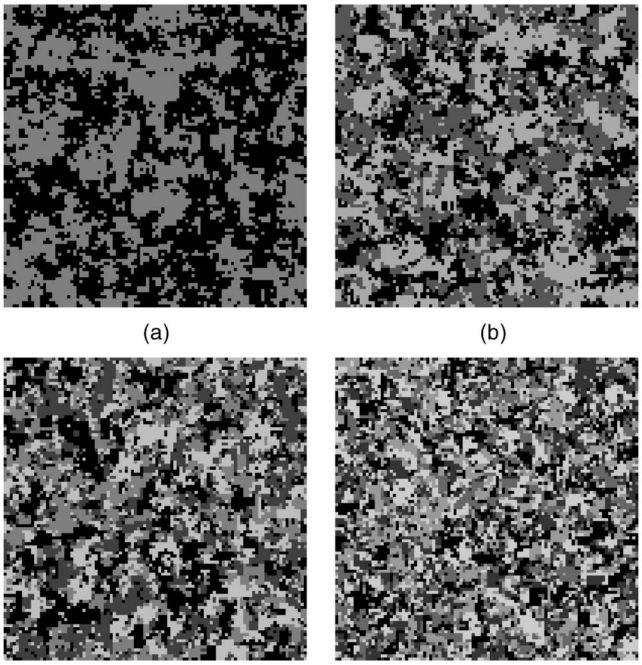


Fig. 1. Simulations of a K-color Potts model (before adding noise) for different values of K and β: (a) $K = 2, \beta = 0.78$, (b) $K = 3, \beta = 0.9$, (c) $K = 4, \beta = 1$, and (d) $K = 5, \beta = 1$.

was used. At each iteration, the simulated field $\tilde{\mathbf{z}}$ was first updated by carrying out only one iteration of the Gibbs sampler for the current parameter value and then one iteration of the EM algorithm was done for the resulting factorized model. As regards estimation of Ψ, so doing the Maximization (M) step in EM becomes tractable. To be more specific, as noted by [13], 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 $\tilde{\mathbf{z}}$. It follows that estimating θ was straightforward since, in this case, a closed-form expression similar to that for finite Gaussian mixtures (see, for instance, [35] chapter 2), is available. Parameter β was also easily obtained through a standard numerical maximization procedure.

Then, $\mathrm{BIC}^{GBF}(\hat{\Psi}_K)$ was computed as defined in (23). We also report values of BIC when the images are seen as realizations of independent mixture models in order to measure the gain of taking spatial information into account when selecting the number of classes. The EM algorithm was used to estimate the parameters and the criterion (computed exactly in this case) is denoted by BIC^{IND}. We also compared with PLIC based on the ICM algorithm as an alternative criterion assuming a spatial model.

When not otherwise specified, the algorithms (Simulated field, EM, and ICM 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. The algorithms were all stopped after $N = 100$ iterations.

The images used for the experiments are described below. In Section 6.1, we first compare the criteria on fully simulated data. The models used for the simulations are the models used in the estimation and segmentation algorithms. In Section 6.2, we consider synthetic images degraded with some simulated Gaussian noise. The true K is known, but the images are not realizations of a known probabilistic model. In Section 6.3, real-life images are considered.

TABLE 1
Degraded K-color Potts Model

$K = 2, \beta = 0.78$	
selected K	2
BIC^{IND}	100
PLIC	100
BIC^{GBF}	100

$K = 3, \beta = 0.9$	
selected K	3
BIC^{IND}	100
PLIC	100
BIC^{GBF}	100

$K = 4, \beta = 1$			
selected K	3	4	5
BIC^{IND}	38	62	0
PLIC	0	100	0
BIC^{GBF}	0	99	1

$K = 5, \beta = 1$			
selected K	4	5	6
BIC^{IND}	79	21	0
PLIC	0	100	0
BIC^{GBF}	0	92	8

$K = 6, \beta = 1.1$				
selected K	4	5	6	7
BIC^{IND}	13	80	7	0
PLIC	0	2	98	0
BIC^{GBF}	0	0	99	1

Selected K using BIC for independent mixture models (BIC^{IND}), pseudolikelihood (PLIC) and mean field-like (BIC^{GBF}) approximations of BIC. The reported values are the number of times a given K is selected out of 100 experiments.

6.1 Hidden K-Color Potts Models

We first tested the criteria on images simulated from hidden Potts models for which the true parameters β and θ were known. We created 100×100 images by simulating 2D K-color Potts models (24) for $K = 2, \ldots 6$ and different values of β using the Gibbs sampler of [31]. We considered a first order neighborhood, i.e., four neighbors for each pixel. We chose β so that the simulated images present homogeneous regions and some spatial structure (e.g., Fig. 1) for in other cases we cannot really expect the criteria to recover the true K. For smaller values of β, typical realizations look much noisier and are visually close to independently distributed colors. For larger values, the simulations lead to close to monocolor images, whatever the true K used for the simulations. Then, a Gaussian noise was added to the Potts model realizations, so that the resulting simulated data are continuously valued images and correspond to hidden K-color Potts models for which $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots K\}$ with $\mu_k = k$ and $\sigma_k = 0.5$, for $k = 1, \ldots K$. We used our knowledge of a constant variance for the K states to fit a model and recover the true image. For each model considered, 100 simulations were carried out. The corresponding criteria results are reported in Table 1. It appears that criteria BIC^{GBF} and PLIC perform well and outperform BIC^{IND}, which shows degradation in selecting the right number of colors when K is larger than 4. This confirms the advantage of using spatial models even through approximations, but does not enable to differentiate BIC^{GBF} from PLIC. PLIC performs slightly better for $K = 4$ and $K = 5$, but the differences cannot be considered as significant. More differences appear in the next two sections.

6.2 Noise-Corrupted Synthetic Images

In this section, we consider noise-corrupted images corresponding to known values of K. Fig. 2b is a 128×128 image

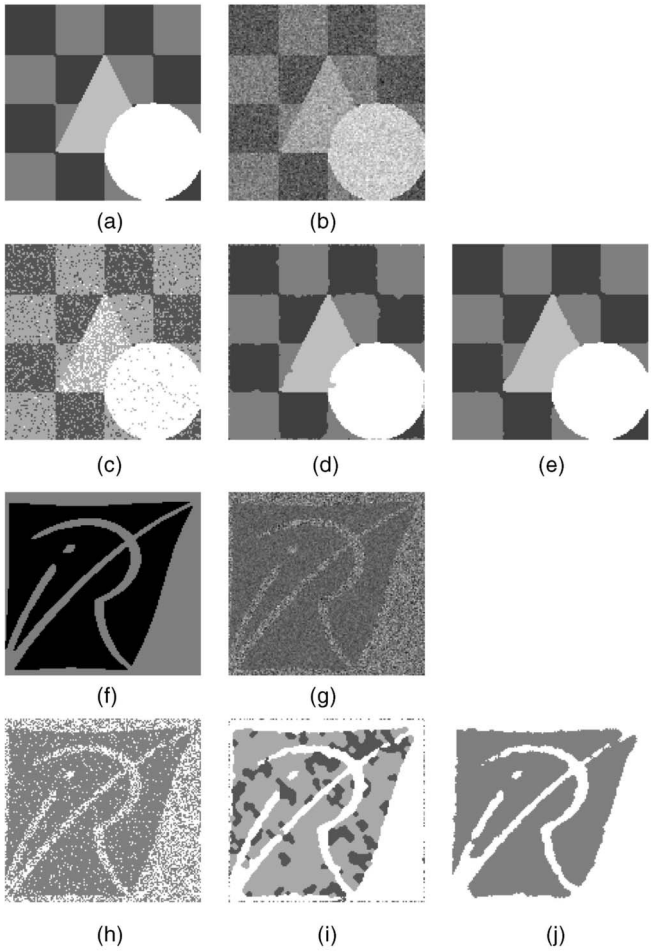


Fig. 2. Noise-corrupted synthetic images. Checkerboard image: (a) original image, (b) noise-corrupted image, (c) 3-color segmentation usingEM for independent mixtures, (d) 4-color segmentation using ICM, and (e) 4-color segmentation using the simulated field algorithm. Logo image: (f) original image, (g) noise-corrupted image, (h) 2-color segmenation using EM for independent mixtures, (i) 3-color segmentation using ICM, and (j) 2-color segmenation using the simulated field algorithm.

obtained by adding some Gaussian noise to the 4-color image of Fig. 2a. The noise parameters are given by $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, 4\}$ with $\mu_k = k$ and $\sigma_k = 0.5$ for $k = 1, \ldots, 4$. The other example (Fig. 2g) is a 133×142 noise-corrupted 2-color image. We used Gaussian densities with class-dependent variances so that the true noise parameters are $(\mu_1, \sigma_1) = (51, 130)$ and $(\mu_2, \sigma_2) = (255, 300)$. These images before degradation are not realizations from a known Markov field model. However, the spatial component is important in these two examples. Using the EM algorithm for independent mixture models to restore the original images, leads to images still noisy (see Figs. 2c and 2h). When considering the selection of K, assuming a nonspatial model can then be inefficient (Table 2). When taking into account the spatial component, we assumed for estimation a model with second order neighborhood (i.e., the eight closest neighbors for each pixel). The selected K are reported in Table 2. In these experiments, BIC^{GBF} and PLIC behave differently. We observe that BIC^{GBF} is better in selecting the right number of colors for images presenting thin features (e.g., Fig. 2f) while they both perform well when images are made of larger

TABLE 2
Noise-Corrupted Synthetic Images

Checkerboard image	
criterion	selected K
BIC^{IND}	3
PLIC	4
BIC^{GBF}	4

Logo image	
criterion	selected K
BIC^{IND}	2
PLIC	3
BIC^{GBF}	2

Selected K using BIC for independent mixture models (BIC^{IND}), pseudolikelihood (PLIC), and mean field-like (BIC^{GBF}) approximations of BIC.

regions (e.g., Fig. 2a). The corresponding segmentations (Figs. 2e and 2j) using the simulated field algorithm are closer to the original images than that obtained using ICM (Figs. 2d and 2i). Additional experiments were carried out with other images containing thin lines and showed similar results in favor of BIC^{GBF}. This may be due to the respective use of the simulated field and ICM algorithms which are of a rather different nature. A well-known feature of ICM is its tendency to produce oversmoothed segmentations while the stochastic nature of the simulated field algorithm makes it more flexible. However, in our experiments, the same difference in the number of classes selected by PLIC and BIC^{GBF} was observed even when the parameter estimations and segmentations were similar for the different values of K. This suggests that the better performance of BIC^{GBF} relies mainly in the way BIC is approximated: a better approximation of the likelihood and a satisfying estimation of the parameters. It is not clear though why this difference with PLIC is not more striking for images made of large monocolor regions. This may come from the nature of BIC itself, but investigating this was out of the scope of this paper. The difficulty of the analysis is increased by the coupling of segmentation and estimation algorithms with model selection criteria.

6.3 Gray-Level Images

We eventually tried the criteria on real images for which a true value for K does not exist (in real-life, it is usually part of the problem to assess its value), but for which intuition or expert knowledge could give an indication of what would be a reasonable value. As an illustration, Fig. 3a is an aerial 100×100 image of a buoy against a background of dark water, and Fig. 3g is a 128×128 PET image of a dog lung (see [19] for more details on their nature and origin).

For the first image, we suspect that 2 is a relevant value for K. Fig. 3a presents some artifact: horizontal scan lines from the imaging process can be observed. Some preprocessing step to remove this known artifact could be carried out as in [19], but we tested here the criteria on the raw data. The selected K are shown in Table 3, and the corresponding segmentations in Fig. 3. BIC^{GBF} performs much better than PLIC, which selects a too large number of components while BIC^{IND} probably suffers from not taking into account the spatial information, as can be seen on Fig. 3d. These results were obtained using basic thresholding to produce initial segmentations for the estimation algorithms (simulated field and ICM algorithms). We tried BIC^{GBF} and PLIC with more refined initializations using the independent mixtures EM algorithm segmentations as first images. This can be seen as a preprocessing step. The selected K was

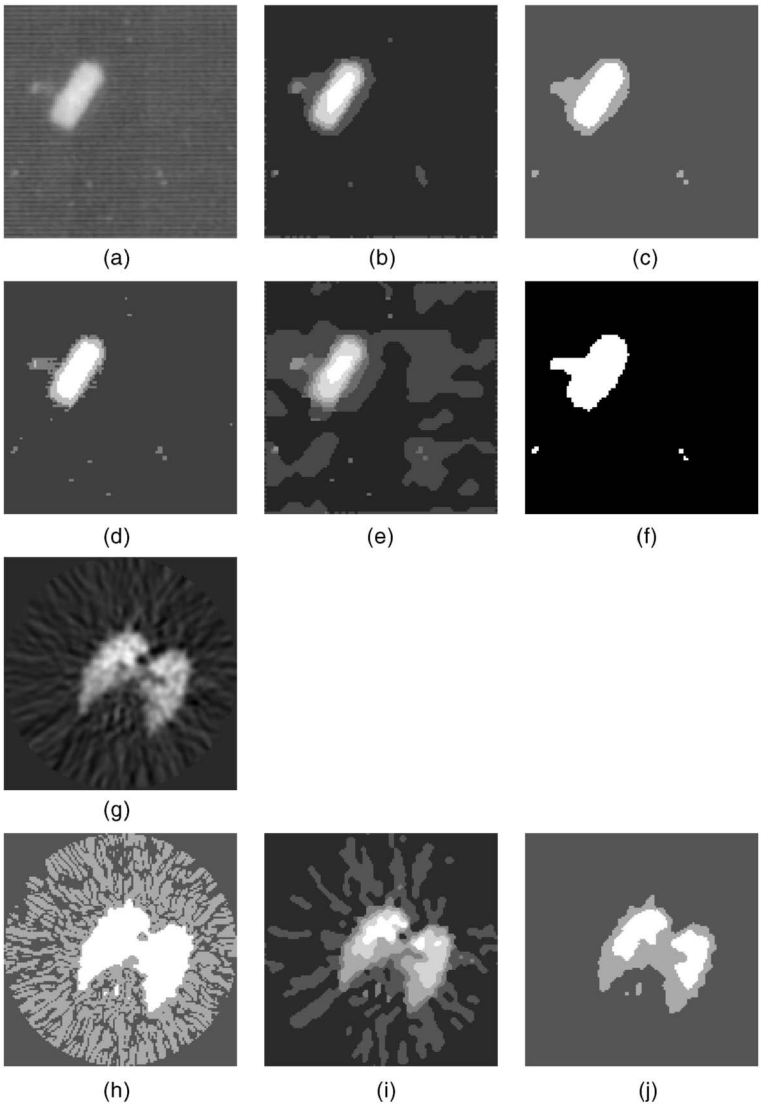


Fig. 3. Gray-level images. Buoy image: (a) original image, (b) and (c) 6 and 3-color segementation using, respectively, ICM and the simulated field algorithm initialized by threshholding, (d) 4-color segmentation using EM for independent mixtures, (e) and (f) 7 and 2-color segmentations using, respectively, ICM and the simulated field algorithm initialized by EM for independent mixtures. PET image of a dog lung: (g) original image, (h) 3-color segmentation using EM for independent mixtures, (i) 6-color segmentation using ICM, and (j) 3-color segmentation using the simulated field algorithm.

then 2 for BIC^{GBF} (Fig. 3f), but still too large (seven classes) for PLIC which leads to a meaningless segmentation (Fig. 3e).

For the dog lung image, the aim is to distinguish the lung from the rest of the image in order to measure the heterogeneity of the tissue in the region of interest. Only pixels in this delimited area will then be considered to compute a heterogeneity measure, such as a coefficient of variation. PLIC and BIC^{GBF} select rather different K with again a too large value for PLIC (Table 3). The corresponding segmentations are shown in Figs. 3i and 3j. The 3-color segmentation obtained using BIC^{GBF} and the simulated field algorithm (Fig. 3j) is the more satisfactory as regards interpretation. It shows one color for the background and two for the lung itself. This is not surprising since the image is constructed based on radioactive emissions from gas in the lung. The two segments account for the high gas density in the interior of the lung and the somewhat lower gas density around the periphery. BIC^{IND} also selects three colors, but the corresponding segmentation is rather different (Fig. 3h), focusing more on the artificial background circle. We then also computed PLIC and BIC^{GBF} using the independent mixtures EM segmentations instead of the ones obtained via

TABLE 3
Gray-Level Images

Buoy image	
criterion	selected K
BIC^{IND}	4
PLIC	6
BIC^{GBF}	3

Dog lung image	
criterion	selected K
BIC^{IND}	3
PLIC	6
BIC^{GBF}	3

Selected K using BIC for independent mixture models (BIC^{IND}), pseudolikelihood (PLIC), and mean field-like (BIC^{GBF}) approximations of BIC.

thresholding as initializing images, but noticed no significant difference.

7 Discussion

In the context of Markov model selection, starting from BIC as our selection criterion, we proposed using mean field-like approximations to deal with the computation of the intractable Markov distribution in BIC expression. More specifically, one of our contributions was to notice that BIC could be rewritten in terms of partition functions for which a first order rather than zeroth order mean field approximation was available (Section 5.2). The advantage is that the quality of the approximation is easier to assess since it uses the best lower bounds for the partition functions. We introduced a class of new criteria among which we chose one, the so-called BIC^{GBF} (23) based on these theoretical considerations regarding the quality of the approximation of the intractable likelihood and based on previous experimental results as regards parameter estimation for various types of images. First, it appears that taking spatial information into account leads to some improvements when compared to BIC for independent mixture models (BIC^{IND}). Then, our criterion differentiates from PLIC (BIC approximation based on the pseudolikelihood) in its ability to deal better with thin features in images. It also shows good performance on real images although we can suspect decreasing performance in the presence of artifact, like scan lines, that the criterion may consider as relevant information instead of noise. However, this is likely to be handled by some preprocessing step using reasonable initializations, as EM for independent mixtures for instance.

After carrying out various experiments, it appeared that a sensible procedure for model selection would be to first perform simple procedures. For example, for selecting the number of components into which to segment an image, a natural procedure is the EM algorithm for independent mixture models easy to implement and for which BIC values can be computed exactly. In some cases, this could lead to reasonably satisfying selection and segmentation so that users may choose not to go further. If not, as it is likely to occur for images with significant spatial structure, the corresponding procedure could possibly be further used to initialize more refined algorithms based on spatial models. For example, [12] studied ICM and used the pseudolikelihood approximation while we propose to use the simulated field algorithm of [15] and the mean field approximation principle to compute criterion BIC^{GBF}.

On the set of images tested in our experiments, our procedure showed much better performance, especially on real data. We believe that this is mainly due to a better approximation of the likelihood in BIC^{GBF} (see the Appendix for an illustration of the superiority of the first order approximation) coupled to a satisfying estimation of the parameter provided by the simulated field algorithm.

This study remains somewhat limited in that it is mainly exploratory and based on experiments. We did not address the question of the consistency of the various criteria. As far as we know, no such results are currently available for hidden Markov random fields. In some recent work, [8] consider a maximized penalized marginal likelihood criterion for estimating the number of hidden states in hidden Markov chains. The author in [8] proves a consistency result for this criterion, although the marginal likelihood involved is not necessarily close to the likelihood (they are equal only when the variables are independent). This suggests that a good approximation of the maximized log-likelihood is not a strong requirement to obtain consistent criteria. A key point in [8] seems to be the decomposition of the criterion as a sum of identically distributed terms. The criteria proposed in this paper can also be written as sum because of the factorization property of the distributions involved. The generalization is not straightforward, but our next step is therefore to investigate if consistency results can be deduced in a similar way.

Appendix

Zeroth and First Order Approximations for the Partition Function of a 2-Color Potts Model

The notation is that of Section 4. Considering simple Potts models, our aim is to illustrate that W^{GBF} (16) can be a better approximation of W than the standard mean field approximation W^{mf}. The energy of a Potts model can be written

$$H(\mathbf{z}|\beta) = -\beta \sum_{i\sim j} z_i^t z_j = -\frac{\beta}{2} \sum_{i=1}^{n} z_i^t \sum_{j\in N(i)} z_j \ ,$$

it follows the zeroth order mean field approximation $H^{mf}(\mathbf{z}|\beta) = -\beta \sum_{i=1}^{n} z_i^t \sum_{j\in N(i)} \overline{z}_j$, with $\overline{z}_j = \mathbb{E}^{mf}[Z_j]$. Then,

$$\mathbb{E}^{mf}[H(\mathbf{Z}|\beta)] = -\frac{\beta}{2} \sum_{i=1}^{n} \overline{z}_i^t \sum_{j\in N(i)} \overline{z}_j = \frac{1}{2}\mathbb{E}^{mf}[H^{mf}(\mathbf{Z}|\beta)],$$

$$\text{so that} \quad W^{mf} = \sum_{\mathbf{z}} \exp(-H^{mf}(\mathbf{z}|\beta))$$

$$= \prod_{i=1}^{n} \sum_{z_i} \exp\left(\beta z_i^t \sum_{j\in N(i)} \overline{z}_j\right),$$

$$\text{and} \quad W^{GBF} = W^{mf} \exp(\mathbb{E}^{mf}[H(\mathbf{Z}|\beta)])$$

$$= W^{mf} \exp\left(-\frac{\beta}{2} \sum_{i=1}^{n} \overline{z}_i^t \sum_{j\in N(i)} \overline{z}_j\right).$$

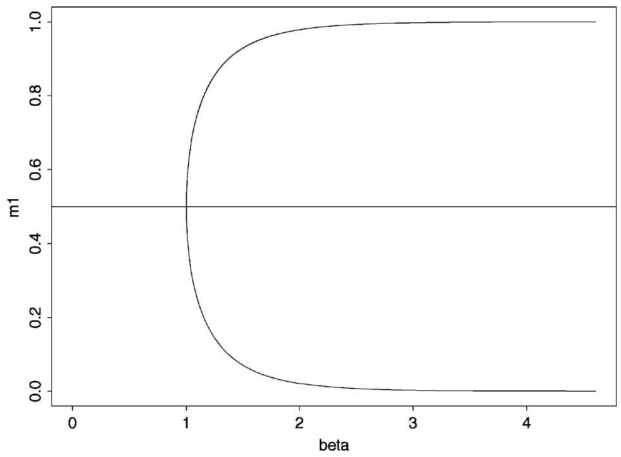
Using symmetries, for all $i = 1, \ldots, n$, we can write $\overline{z}_i = \mathrm{m}$ with m being, in the two-color case, the two-component vector $(m_1, m_2)^t$ satisfying $m_1 + m_2 = 1$ and the following consistency conditions,

$$m_1 = \frac{\exp(\beta N m_1)}{\exp(\beta N m_1) + \exp(\beta N m_2)}$$
$$m_2 = \frac{\exp(\beta N m_2)}{\exp(\beta N m_1) + \exp(\beta N m_2)},$$

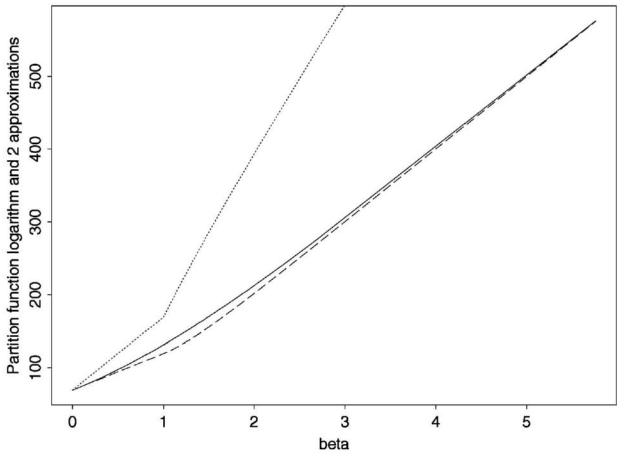
where $N = |N(i)|$ is the number of neighbors assumed the same for all sites.

This is equivalent to solve

$$m_1 = \frac{\exp(\beta N m_1)}{\exp(\beta N m_1) + \exp(\beta N (1 - m_1))}$$
$$= \frac{1}{1 + \exp(\beta N (1 - 2m_1))}.$$



(a)



(b)

Fig. 4. One-dimensional ($\mathrm{N} = 2$ neighbors) 2-color Potts model: (a) solutions $(m_1, 1-m_1)$ of the mean field consistency conditions as β varies, (b) for $n = 100$ sites, exact partition function logarithm and two approximations for $\beta > 0$. Solid line shows the exact $\log W$, wider dashed line shows $\log W^{GBF}$ and smaller dashed line shows $\log W^{mf}$.

Note that, if m_1 satisfies (25), then $1-m_1$ is also a solution. For $\beta < K/N$, i.e., $\beta < 2/N$ there is only one solution $m_1 = 1/2$. For $\beta > 2/N$, there are two additional solutions m_1 and $1-m_1$ with $m_1 > 1/2$. We focus on solutions $m_1 \neq 1/2$. Such a solution is a nonconstant function of β whose closed form expression is not available. However, using (25), β can be expressed as a function f of m_1 given by

$$\beta = f(m_1) \;=\; \frac{1}{N(1-2m_1)} \log\left(\frac{1-m_1}{m_1}\right). \tag{26}$$

It is easy to check that $f(1-m_1) = f(m_1)$ so that two symmetric solutions lead to the same β as expected. We can also check that $f(m_1)$ tends to $2/N$ when m_1 tends to $1/2$ and to infinity when m_1 tends to 1. The graph of m_1 against β is shown in Fig. 4a.

The quantity m_1 appears in the expressions of W^{mf} and W^{GBF}, while the true W depends only on β. However, when W is available in closed form, using (26), the three quantities can be expressed and compared as functions of m_1. For periodic boundary conditions, it comes

$$W^{mf} = (\exp(\beta N m_1) + \exp(\beta N(1-m_1)))^n \tag{27}$$

$$W^{GBF} = W^{mf} \exp(-\frac{\beta}{2} N n (m_1^2 + (1-m_1)^2)). \tag{28}$$

It follows, using (25)

$$\begin{aligned} \log(W^{mf}) &= \beta N n m_1 + n \log(1 + \exp(\beta N(1-2m_1))) \\ &= \beta N n m_1 - n \log(m_1) \end{aligned} \tag{29}$$

$$\begin{aligned} \text{and, } \log(W^{GBF}) = &\frac{\beta}{2} N n (4m_1 - 2m_1^2 - 1) \\ &+ n \log(1 + \exp(\beta N(1-2m_1))). \end{aligned} \tag{30}$$

As regards W, a closed form is not available, in general. However, in the 1-dimensional case for which $N = 2$, an expression of W is $W = (\exp(\beta)+1)^n + (\exp(\beta)-1)^n$. It is then easy to compare the logarithms. For $N = 2$,

$$\begin{aligned} \log(W^{mf}) &= 2nm_1\beta + n\log(1+\exp(2\beta(1-2m_1))) \\ \log(W^{GBF}) &= n(4m_1 - 2m_1^2 - 1)\beta \\ &\quad + n\log(1+\exp(2\beta(1-2m_1))) \\ \log(W) &= n\beta + n\log(1+\exp(-\beta)) \\ &\quad + \log\left(1 + \left(\frac{1-\exp(-\beta)}{1+\exp(-\beta)}\right)^n\right). \end{aligned}$$

For $\beta < 1$, $m_1 = 1/2$, it comes

$$\begin{aligned} \log(W^{mf}) &= n\beta + n\log(2), \\ \log(W^{GBF}) &= n\frac{\beta}{2} + n\log(2). \end{aligned}$$

The corresponding graphs are shown in Fig. 4b.

When $\beta > 1$, there are no analytical expressions for m_1 as a function of β, but we can plot the graphs by inverting (26) (See Fig. 4b). Note that $\log(W^{mf})$ and $\log(W^{GBF})$ remain the same when m_1 is changed to $1-m_1$. It appears clearly on the plot that $\log(W^{GBF})$ is a far better approximation of the exact $\log(W)$ than $\log(W^{mf})$.

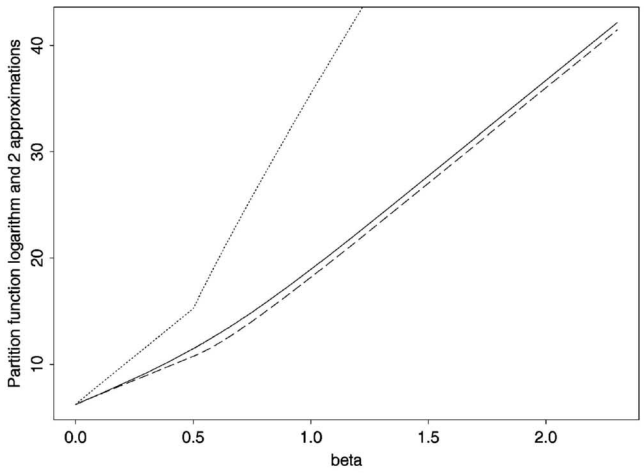
For dimension greater than 1, the mean field approximation expressions (27) and (28) are still valid, but the computation of the true W is exponentially complex. We are restricted then to a 3×3 grid, i.e., $n = 9$ sites and considered successively $N = 4$ and $N = 8$ neighbors. For $N = 4$, the exact partition function is,

$$\begin{aligned} W \;= &102\exp(6\beta) + 144\exp(8\beta) + 198\exp(10\beta) + \\ &48\exp(12\beta) + 18\exp(14\beta) + 2\exp(18\beta). \end{aligned}$$

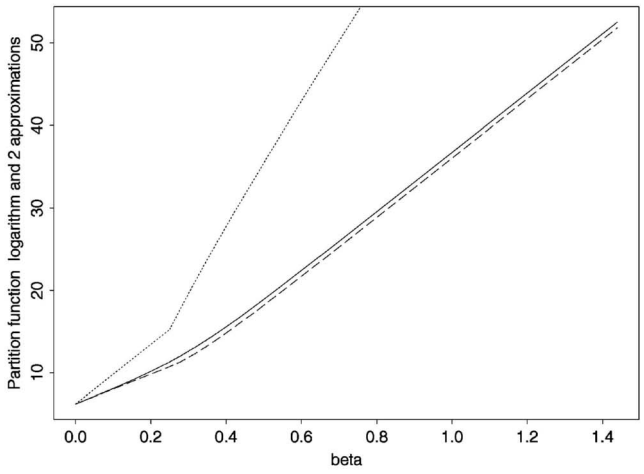
For $N = 8$, it comes

$$\begin{aligned} W \;= &252\exp(16\beta) + 168\exp(18\beta) + 72\exp(22\beta) + \\ &18\exp(28\beta) + 2\exp(36\beta). \end{aligned}$$

The partition function logarithm and its approximations are shown in Fig. 5. In the general case, when β tends to infinity, W behaves (if K denotes the number of colors) as $K\exp(nN\beta/2)$, which is the dominant term in the sum over all possible configurations. The term $nN/2$ is the maximum number of homogeneous cliques. It occurs for each of the K monocolor configurations. Therefore, when β tends to



(a)



(b)

Fig. 5. Two-color Potts model on a 3×3 grid with (a) $\mathrm{N} = 4$ neighbors, (b) $\mathrm{N} = 8$ neighbors: exact partition function logarithm and two approximations for $\beta > 0$. Solid line shows the exact $\log W$, wider dashed line shows $\log W^{GBF}$, and smaller dashed line shows $\log W^{mf}$.

infinity, $\log(W)$ behaves as $nN\beta/2 + \log K$. When looking at (29) and (30), it appears that when β tends to infinity, m_1 tends to 0 or 1 and, in both cases, $\log(W^{mf})$ behaves as $nN\beta$ and $\log(W^{GBF})$ as $nN\beta/2$. This suggests ways to improve the approximations. The $\log(2) = 0.69$ difference between $\log(W)$ and $\log(W^{GBF})$ appears more clearly in Fig. 5. Again, $\log(W^{GBF})$ appears to be a much better approximation of $\log(W)$ than $\log(W^{mf})$.

ACKNOWLEDGMENTS

The authors are grateful to G. Celeux for many valuable comments.

REFERENCES

[1] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics,* vol. 6, pp. 461-464, 1978.
[2] M. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," *Proc. Second Int'l Symp. Information Theory,* B.N. Petrox and F. Caski, eds., pp. 267-281, 1973.
[3] J. Rissanen, "Stochastic Complexity in Statistical Inquiry," *Word Scientific,* 1989.
[4] P. Zhang, "Model Selection via Multifold Cross Validation," *The Annals of Statistics,* vol. 21, pp. 299-313, 1993.
[5] R. Kass and A. Raftery, "Bayes Factor," *J. Am. Statistical Assoc.,* vol. 90, pp. 733-795, 1995.
[6] J.O. Berger and T. Sellke, "Testing a Point Null Hypothesis: The Irreconcilability of P-Values and Evidence," *J. Am. Statistical Assoc.,* vol. 82, pp. 112-122, 1987.
[7] A. Raftery, "Bayesian Model Selection in Social Research (with discussion)," *Sociological Methodology,* P.V. Marsden, ed., Cambridge, Mass.: Blackwell, pp. 111-163, 1995.
[8] E. Gassiat, "Likelihood Ratio Inequalities with Applications to Various Mixtures," Technical Report 2001-20, Mathematiques, Orsay, 2001.
[9] C. Ji and L. Seymour, "A Consistent Model Selection Procedure for Markov Random Fields Based on Penalized Pseudolikelihood," *Annals of Applied Probability,* vol. 6, pp. 423-443, 1996.
[10] J. Besag, "Statistical Analysis of Non-Lattice Data," *The Statistician,* vol. 24, pp. 179-195, 1975.
[11] L. Seymour and C. Ji, "Approximate Bayes Model Selection Procedures for Gibbs-Markov Random Fields," *J. Statistical Planning and Inference,* vol. 51, pp. 75-97, 1996.
[12] D. Stanford and A. E. Raftery, "Determining the Number of Colors or Gray Levels in an Image Using Approximate Bayes Factors: The Pseudolikelihood Information Criterion (PLIC)," technical report, Dept. of Statistics, Univ. of Washington, http://www.stat.washington.edu/, Feb. 2001.
[13] W. Qian and D.M. Titterington, "Estimation of Parameters in Hidden Markov Models," *Philosophical Trans. Royal Soc. London A,* vol. 337, pp. 407-428, 1991.
[14] D. Chandler, *Introduction to Modern Statistical Mechanics.* Oxford Univ. Press, 1987.
[15] G. Celeux, F. Forbes, and N. Peyrard, "EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation," *Pattern Recognition,* vol. 36, no. 1, pp. 131-144, 2003.
[16] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, pp. 719-725, 2000.
[17] J.M. Hammersley and P.E. Clifford, "Markov Fields on Finite Graphs and Lattices." 1971.
[18] D. Geman, "Random Fields and Inverse Problems in Imaging," *Lecture Notes in Math.,* vol. 1427, New York: Springer, pp. 113-193, 1991.
[19] D. Stanford, *Fast Automatic Unsupervised Image Segmentation and Curve Detection in Spatial Point Processes.* PhD thesis, Dept. of Statistics, Univ. of Washington, Seattle, 1999.
[20] G.J. McLachlan and D. Peel, *Finite Mixture Models.* Wiley, 2000.
[21] C. Fraley and A. Raftery, "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *Computer J.,* vol. 41, pp. 578-588, 1998.
[22] K. Roeder and L.A. Wasserman, "Practical Bayesian Density Estimation Using Mixtures of Normals," *J. Am. Statistical Assoc.,* vol. 92, pp. 894-902, 1997.
[23] M. Newton and A. Raftery, "Approximate Bayesian Inference by the Weighted Likelihood Bootstrap (with discussion)," *J. Royal Statistical Soc. B,* vol. 56, pp. 3-48, 1994.
[24] D. Geiger and F. Girosi, "Parallel and Deterministic Algorithms from MRFs: Surface Reconstruction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 5, pp. 401-412, May 1991.
[25] J. Zerubia and R. Chellappa, "Mean Field Approximation Using Compound Gauss-Markov Random Field for Edge Detection and Image Restoration," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 2193-2196, 1990.
[26] A.L. Yuille, "Generalized Deformable Models, Statistical Physics and Matching Problems," *Neural Computation,* vol. 2, pp. 1-24, 1990.
[27] T.S. Jaakkola and M.I. Jordan, "Improving the Mean Field Approximation via the Use of Mixture Distributions," *Learning in Graphical Models,* M.I. Jordan, ed., Dordrencht: Kluwer Academic Publishers, pp. 163-173, 1998.
[28] T. Hofmann and M. Buhmann, "Pairwise Data Clustering by Deterministic Annealing," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 1, pp. 1-14, Jan. 1997.

[29] G.E.B. Archer and D.M. Titterington, "Parameter Estimation for Hidden Markov Chains," *J. Statistical Planning Inference,* 2002.
[30] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. Royal Statistical Soc. B,* vol. 48, pp. 259-302, 1986.
[31] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, pp. 721-741, 1984.
[32] N. Peyrard, "Approximations de Type Champ Moyen des Modèles de Champ de Markov pour la Segmentation de Données Spatiales," PhD thesis, U.F.R. d'Informatique et de Math. Appliquées, Univ. Joseph Fourier, Grenoble I, France, 2001.
[33] Z. Zhou, R. Leahy, and J. Qi, "Approximate Maximum Likelihood Hyperparameter Estimation for Gibbs Priors," *IEEE Trans. Image Processing,* vol. 6, no. 6, pp. 844-861, 1997.
[34] G. Potaniamos and J. Goutsias, "Stochastic Approximation Algorithms for Partition Function Estimation of Gibbs Random Fields," *IEEE Trans. Information Theory,* vol. 43, no. 6, pp. 1948-1965, 1997.
[35] G.J. McLachlan and K.E. Basford, *Mixture Models: Inference and Applications to Clustering.* Dekker, 1987.

Florence Forbes received the PhD degree in applied probabilities in 1996, from University Joseph Fourier, Grenoble, France. She is a research scientist at the Institut National de Recherche en Informatique et Automatique (INRIA). 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.

Nathalie Peyrard received the PhD degree in statistics in 2001, from University Joseph Fourier, Grenoble. She is currently a member of the VISTA research team, doing a postdoctoral at Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) in Rennes, France. Her research interests include spatial statistics, Markov random fields, stochastic algorithms (MCMC), and applications in image analysis.