Bayesian inference on a mixture model with spatial dependence

Lionel Cucala and Jean-Michel Marin
I3M UMR CNRS 5149, Université Montpellier 2, France
May 6, 2012

Abstract

We introduce a new technique to select the number of components of a mixture model with spatial dependence. It consists in an estimation of the Integrated Completed Likelihood based on a Laplace’s approximation and a new technique to deal with the normalizing constant intractability of the hidden Potts model. Our proposal is applied to a real satellite image.

Keywords: Hidden Gibbs random fields, Bayesian model choice, Integrated Completed Likelihood, auxiliary Markov Chain Monte Carlo schemes

1 Introduction

We consider a finite set of sites $S = \{1, \cdots, n\}$. At each site $i \in S$, we observe $z_i \in \mathcal{X}_i$ where $\mathcal{X}_i$ is a finite set of states. We also consider a graph $\mathcal{G}$ which defines neighborhood relations: sites $i$ and $i'$ are neighbors if there is an edge between $i$ and $i'$ in the graph $\mathcal{G}$. For a given graph, a clique $c$ is a subset of $S$ where all elements are mutual neighbors. We note $\mathcal{C}$ the set of cliques associated to $\mathcal{G}$. A Gibbs random field (Cressie, 1993; Rue and Held, 2005) is a probabilistic model whose the probability density function with respect to the counting measure is given by

$$f(z) = \frac{1}{Z} \exp\{-U(z)\} = \frac{1}{Z} \exp\left\{- \sum_{c \in \mathcal{C}} U_c(z) \right\},$$

where $U(z) = \sum_{c \in \mathcal{C}} U_c(z)$ is called the potential function and $Z$ is the normalizing constant such that

$$Z = \sum_{z \in \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n} \exp\left\{- \sum_{c \in \mathcal{C}} U_c(z) \right\}.$$

We are interested in Gibbs random fields admitting a potential function $U(z) = -\beta^T S(z)$ where $\beta \in \mathbb{R}^p$ is a scale parameter and $S(\cdot)$ is a function taking values in $\mathbb{R}^p$ such that $S(z) = \sum_{c \in \mathcal{C}} S_c(z)$. 
In that case,

\[ f(z|\beta) = \frac{1}{Z_\beta} \exp\{\beta^T S(z)\} \]

and the normalizing constant \( Z_\beta \) depends on the scale parameter \( \beta \). These models are used to represent spatial dependence with numerous applications notably in image analysis (Hurn et al., 2003; Alfò et al., 2008), disease mapping (Green and Richardson, 2002), genetic analysis (François et al., 2006) among others. If a Markov random field admits a probability density function which is everywhere positive, then the Hammersley-Clifford theorem establishes that there exists a corresponding Gibbs random field representation. That is the case of the Potts model for which

\[ S(z) = \sum_{i \sim i'} \mathbb{I}\{z_i = z_{i'}\} \]

where \( \sum_{i \sim i} \) means that the summation is taken over all the pairs of neighbor pixels.

In the most realistic cases, the summation

\[ Z_\beta = \sum_{z \in X_1 \otimes \ldots \otimes X_n} \exp\{\beta^T S(z)\} \]

implies too many terms. For a given model, the fact that the calculation of the likelihood is intractable makes very difficult the inference process on \( \beta \). The difficulties are more important when one wishes to compare models. In such a case, we would like to determine the best Gibbs random fields within a collection of \( M \) models. Each model is associated to a sufficient statistics \( S_m \) (1 \( \leq \) \( m \) \( \leq \) \( M \)) and the following probability density function

\[ f_m(z|\beta_m) = \exp \left\{ \beta_m^T S_m(z) \right\} / Z_{\beta_m,m} \]

where \( \beta_m \in \Theta_m \) and \( Z_{\beta_m,m} \) is the unknown normalizing constant. When \( z \) is observed, the most recent Bayesian procedure is the ABC approach of Grelaud et al. (2009). That one is quite easy to implement, gives very good results but the computing time involved can be very important.

Here, we consider that the Gibbs random field is not directly observed, like for Hidden Markov Models (HMMs). Only a signal \( y \) linked to \( z \) by the conditional probability distribution \( g(y|z, \lambda) \) parametrized by \( \lambda \) is observed. In such a case, we would like to estimate the parameters \( \beta_m \) and \( \lambda \), but also to select a Gibbs random field. There exists very few works on this model choice problem. The most relevant one belongs to the frequentist paradigm. That is the proposal of Forbes and Peyrard (2003) based on mean-field approximations.
The goal of this paper is to propose a way to select an hidden Gibbs random field. We focus on Bayesian approaches and concentrate on the case of Gaussian mixture models with spatial dependence. The hidden field is a Potts model. In Section 2, we give a short recap on such models. In Section 3, we propose a procedure to estimate the parameters of the Potts and mixture models when the number of components is fixed. The number of components selection difficulties are addressed in Section 4. Then, we apply our proposal to a real satellite image of the Menteith Lake in the last Section.

2 Gaussian mixture models with spatial dependence

We consider an $N = n^2$ pixels image. $\mathbf{y} = (y_1, \cdots, y_N) \in \{1, \cdots, 256\}^\otimes N$ is the field of the grey-scale values observed on any pixel. The goal is to cluster the pixels into $k$ groups in order to detect some homogenous regions. In Section 5, our approach is applied to a $100 \times 100$ satellite image ($N = 10000$) representing the area of Menteith Lake, in Scotland, and illustrated by Figure 1.

We introduce the hidden field $\mathbf{z}_k = (z_{1,k}, \cdots, z_{N,k}) \in \mathcal{Z}_k = \{1, \cdots, k\}^\otimes N$ representing the component of any pixel. The grey-scale values are Gaussian random variables depending on the component of the associated pixel, so that

$$f(y_i | z_{i,k} = j, \theta_k) = \frac{1}{\sqrt{2\pi\sigma_k}} \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu_{j,k}}{\sigma_k} \right)^2 \right].$$

Figure 1: Satellite image of the Menteith Lake.
Let $\theta_k = (\mu_{1,k}, \cdots, \mu_{k,k}, \sigma_k, \beta_k) \in \mathbb{R}^{k+2}$ be the parameters vector. The variances are supposed to be equal but this assumption can easily be relaxed.

A classical hypothesis would be considering that the $z_{i,k}$'s are independent. In that case, we would deal with a classical Gaussian mixture model. However, we believe that a neighboring information has to be included through the use of a Gibbs random field for the component variable.

We suppose that $z_k$ is distributed according to a Potts model with a unique scale parameter:

$$f(z_k | \beta_k) = \exp \left[ \beta_k \sum_{l \sim i} 1 \left( z_{i,k} = z_{l,k} \right) \right] / Z(\beta_k),$$

where

$$Z(\beta_k) = \sum_{z_k \in \mathbb{Z}_k} \exp \left[ \beta_k \sum_{l \sim i} 1 \left( z_{i,k} = z_{l,k} \right) \right]$$

is the normalizing constant and $S(z_k) = \sum_{l \sim i} 1 \left( z_{i,k} = z_{l,k} \right)$ the sufficient statistic. For the underlying graph, we assume a standard four-neighbor structure, such that the number of pairs of neighbor pixels is $N_{nei} = 2n(n - 1)$.

The joint distribution is thus

$$f(y, z_k | \theta_k) = \prod_{i=1}^N f(y_i | z_{i,k}, \theta_k) f(z_k | \beta_k)$$

and the distribution of the observations is

$$f(y | \theta_k) = \sum_{z_k \in \mathbb{Z}_k} f(z_k | \beta_k) \prod_{i=1}^N f(y_i | z_{i,k}, \theta_k).$$

When one only observes the field $y$, doing inference on such a model is a tedious problem. In the next Section, we show how it can be solved in the Bayesian setting.

3 Estimation of $\theta_k$ when $k$ fixed

The most standard approach to the estimation of $\beta_k$ is to use the pseudo-likelihood approximation following among others Besag et al. (1991), Heikkinen and Hogmander (1994), Rydén and Titterington (1998), Huang and Ogata (2002) and Dryden et al. (2002)...

Unfortunately, as demonstrated in Cucala et al. (2009), this crude technique can lead to very poor approximations. In the same way, Friel et al. (2009) present approximate methods to calculate the likelihood for large lattices based on exact methods for smaller lattices and implement their approach for a small auto-logistic model. Some authors propose to estimate the normalizing constant and then use the obtained
approximation. One can for instance use the path sampling approach (Gelman and Meng, 1998). For not binary large image, the computing time involved is very considerable. Moreover, McGrory et al. (2009) and McGrory and Titterington (2009) show how a variational Bayes scheme can be constructed to analyze a hidden Potts model and demonstrate the results on binary images. In a recent paper Everitt (2012) proposes a ABC procedure. In the frequentist paradigm, Celeux et al. (2003) introduce EM procedures using mean-field like approximations.

Here, we consider the Bayesian setting and use the following prior distributions:

- \( \sigma_k^2 \sim IG(a, b) \),
- \( \mu_k = (\mu_{1,k}, \ldots, \mu_{k,k}) \sim U(0 \leq \mu_{1,k} \leq \cdots \leq \mu_{k,k} \leq 256) \),
- \( \beta_k \sim U([0, 3]) \).

The values of the hyper-parameters \( a \) and \( b \) are chosen such that the Inverse Gamma distribution has heavy tails. We do not discuss here the choice of the prior distributions. You can easily modify the previous statement. The posterior distribution of \( \theta_k \) is such that:

\[
\pi(\theta_k | y) \propto \sum_{z_k \in Z_k} \left\{ \exp \left[ \beta_k \sum_{l \sim i} \mathbb{I}(z_{i,k} = z_{l,k}) \right] / Z(\beta_k) \right\} \prod_{i=1}^N f(y_i | z_{i,k}, \theta_k) \pi(\theta_k).
\]

Recently, Møller et al. (2006) propose an auxiliary variable MCMC algorithm for simulating from the distribution \( \pi(\beta_k | z_k) \) for which the troublesome constant is cancelled out. This work is developed further in Murray et al. (2006). As demonstrated in Friel and Pettitt (2011) and Everitt (2012), the exchange algorithm of Murray et al. (2006) produces very accurate results. Following the idea of Rydén and Titterington (1998), we suggest to use this algorithm within a Gibbs sampler. That consists to iterate calls to the following full conditional distributions:

- \( \sigma_k^2 \sim IG \left( \frac{N}{2} + a, \frac{\sum_{i=1}^N (y_i - \mu_{z_i,k,k})^2}{2} + b \right) \);
- \( \mu_{j,k} \sim N(\bar{y}_{j,k}; \frac{\sigma_k^2}{n_{j,k}}) \mathbb{I}(\mu_{j-1,k} \leq \mu_{j,k} \leq \mu_{j+1,k}), \quad \forall j = 1, \ldots, k \) \( n_{j,k} = \sum_{i=1}^N z_{i,k} = j \);
- \( \beta_k \) is updated using the scheme of Murray et al. (2006);
- \( z_{i,k} \sim \mathcal{M}(1; \omega_{i,1,k}, \ldots, \omega_{i,k,k}), \quad \forall i = 1, \ldots, N \) with

\[
\omega_{i,j,k} = \frac{\exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu_{j,k}}{\sigma_k} \right)^2 + \beta_k \sum_{l \sim i} \mathbb{I}(z_{l,k} = j) \right]}{\sum_{j=1}^k \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu_{j,k}}{\sigma_k} \right)^2 + \beta_k \sum_{l \sim i} \mathbb{I}(z_{l,k} = j) \right]}.
\]
The Murray et al. (2006) algorithm allows us to simulate $\beta_k$ according to the distribution $\pi(\beta_k|z_k)$. The idea is to simulate from an augmented space distribution which take into account two auxiliary variables $\beta'_k$ and $z'_k$:

$$
\pi(\beta_k, \beta'_k, z'_k) \propto f(z_k|\beta_k)\pi(\beta_k)Q(\beta'_k|\beta_k)f(z'_k|\beta'_k).
$$

Given $\beta'_k$, the distribution of $z'_k$ belongs to the same parametric family as $z_k$ (the Potts model). The distribution $Q(\beta'_k|\beta_k)$ is any distribution for the augmented variable $\beta'_k$ which might depends on $\beta_k$. It is clear that the marginal distribution of $\beta_k$ corresponding to (1) is the distribution of interest. Recall that the prior distribution on $\beta_k$ is the uniform distribution on $[0, 3]$ and $Q(\beta'_k|\beta_k)$ is chosen to be the uniform distribution on $[9\beta_k/10; 11\beta_k/10] \cap [0, 3]$. The Murray scheme is a Metropolis-within-Gibbs algorithm targeted for (1) that works as follows:

- Gibbs update of $\beta'_k$ and $z'_k$: $\beta'_k \sim Q(.|\beta^{(t)}_k)$ and $z'_k \sim f(.|\beta'_k)$;
- Exchange move: $\beta^{(t+1)}_k = \beta'_k$ with probability $\min(1, p^{(t)})$, else $\beta^{(t+1)}_k = \beta^{(t)}_k$:

$$
p^{(t)} = \frac{f(z_k|\beta'_k)\pi(\beta'_k)Q(\beta'_k|\beta^{(t)}_k)f(z'_k|\beta^{(t)}_k)}{f(z_k|\beta^{(t)}_k)\pi(\beta^{(t)}_k)Q(\beta^{(t)}_k|\beta'_k)f(z'_k|\beta'_k)}
= \min(11\beta'_k/10, 3) - 9\beta'_k/10
\min(11\beta^{(t)}_k/10, 3) - 9\beta^{(t)}_k/10
\exp\left[\left(\beta^{(t)}_k - \beta'_k\right)\left(\sum_{i\sim j} 1(z'_{i,k} = z'_{j,k}) - \sum_{i\sim j} 1(z_{i,k} = z_{j,k})\right)\right].
$$

We use a Swendsen-Wang algorithm to simulate from the Potts model, as this scheme mixes more quickly than a classical Gibbs sampler. Only few iterations are necessary (Everitt, 2012).

4 Choosing the number of components

For a uniform prior in the space of models, the bayesian paradigm to choose an appropriate number of components consists in maximizing the integrated likelihood. Indeed, if $\mathcal{M}$ is the random value representing the number of components, the Bayesian model choice is done according to

$$
\mathbb{P}(\mathcal{M} = k|y) \propto \mathbb{P}(\mathcal{M} = k) \int f(y|\theta_k)\pi(\theta_k)d\theta_k
\propto \mathbb{P}(\mathcal{M} = k) m_k(y)
$$
where

\[
m_k(y) = \int \sum_{z_k \in Z_k} f(z_k | \beta_k) \prod_{i=1}^{N} f(y_i | z_{i,k}, \theta_k) d\theta_k
\]

= \int \sum_{z_k \in Z_k} \frac{\exp \left[ \beta_k \sum_{l \sim i} \mathbb{1}(z_i, k = z_l, k) \right]}{Z(\beta_k)} \prod_{i=1}^{N} f(y_i | z_{i,k}, \theta_k) d\theta_k.

As this is practically unfeasible, we need to use an approximation of this expression. Contrary to the case of observed Gibbs random fields, we cannot easily implement a ABC model choice procedure: we could not find any good summary statistics and the use of the whole dataset is impossible. Contrary to the case of standard mixture models, we failed to implement efficiently the Chib methodology as the computing time involved is very considerable. Currently, most people use approximations of BIC (which is itself an asymptotic approximation of the evidence) such as the mean-field approximation (Forbes and Peyrard, 2003) but the question we may ask is: is the evidence (or its BIC approximation) a clever criterion?

For standard mixture models, Biernacki et al. (2000) argued that the goal of clustering is not the same as that of estimating the best approximating mixture model, and so BIC may not be the best way of determining the number of clusters. Instead they proposed the Integrated Completed Likelihood (ICL) criterion, whose purpose is to assess the number of mixture components that leads to the best clustering and not to the best fit. This criterion satisfies the following equality:

\[\text{ICL} = \text{BIC} + \text{Entropy of the corresponding clustering}\]

and the number of clusters favored by ICL tends to be smaller than the number favored by BIC because of the additional entropy term. The ICL criterion has now become the gold standard to select the number of clusters of a mixture model.

In order to approximate the integrated completed likelihood

\[ICL(k) = \int f(y, z_k | \theta_k) \pi(\theta_k) d\theta_k\]

Biernacki et al. (2000) proposed using a Laplace approximation, leading to the criterion

\[-2 \log f(y, z_k^* | \theta_k^*) \pi(\theta_k^*) + (k + 2) \log n,\]

where the values \(\theta_k^*\) and \(z_k^*\) satisfy

\[f(y, z_k^* | \theta_k^*) \pi(\theta_k^*) = \max_{z_k, \theta_k} f(y, z_k | \theta_k) \pi(\theta_k)\]
and $k + 2$ is the number of parameters in the model with $k$ components. The missing data have been replaced by their most probable values: the clustering that we will use.

We may write

$$f(y, z_k|\theta_k) = \prod_{i=1}^{N} f(y_i|z_{i,k}, \theta_k)f(z_k|\beta_k)$$

but the last term is intractable since it contains the unknown normalizing constant $Z(\beta_k)$.

A first solution to estimate the Potts model density $f(z_k|\beta_k)$ would be to use the path sampling method. This has already been used in the same context by Marin and Robert (2007), but only for one value of $k$. Since this technique is very time-consuming, applying it for different values of $k$ seems impossible.

A second solution would be to replace the likelihood $f(z_k|\beta_k)$ by the pseudo-likelihood (Besag, 1974)

$$\tilde{f}(z_k|\beta_k) = \prod_{i=1}^{N} \frac{\exp[\beta_k \sum_{l\sim i} 1(z_{l,k} = z_{i,k})]}{\sum_{j=1}^{k} \exp[\beta_k \sum_{l\sim i} 1(z_{l,k} = j)]}.$$ 

It has been shown that the values of the parameter $\beta_k$ maximizing the likelihood and the pseudo-likelihood are very close, specially when dealing with a large number of pixels. However, this is not enough to consider the pseudo-likelihood as a good estimator of the likelihood.

The solution we propose is quite different. Using the Swendsen-Wang algorithm, we can simulate a relative high number of realisations of the Potts model in a short period of time: this allows to estimate the conditional distribution $f(S(z_k)|\beta_k)$ for any value of $\beta_k$, using for example a kernel estimate (dimension 2). The bandwidths are selected according to the criterion of Bashtannyk and Hyndman (2001).

Then we may use the following relation:

$$f(z_k|\beta_k) = \frac{f(S(z_k)|\beta_k)}{N(S(z_k))}$$

where $N(s) = \sum_{z_k} 1(S(z_k) = s)$ is the number of fields such that the sufficient statistic is $s$. We get

$$N(s) = k^N P_{\beta_k=0}(S(z_k) = s)$$

where $k^N$ is the total number of fields and $P_{\beta_k=0}(.)$ is the probability when the components of all pixels are independently and uniformly distributed on $\{1, \cdots, k\}$.

A naive method for estimating $P_{\beta_k=0}(S(z_k) = s)$ consists in simulating fields without spatial dependence and computing the sufficient statistic $S(z_k)$ for each. Figure 2 illustrates this method a 100 $\times$ 100 image and a four-neighbor structure. We simulated 1,000,000 fields without spatial dependence.
Figure 2: Values of \( S(z_k) \) obtained when \( \beta_k = 0 \) on a 100×100 image and a four-neighbor structure.

The range of the obtained values is very small compared to the theoretical one \([0, 19800]\). With this method, we cannot estimate \( \mathbb{P}_{\beta_k=0}(S(z_k) = s) \) for high values of \( s \) such that the ones we are focused on.

Let \( v_{i,j}, i \sim j \), denote the value of the vertex \((i, j)\), i.e. the difference of components between neighbor pixels \( i \) and \( j \):

\[
v_{i,j} = (z_{j,k} - z_{i,k}) \mod k.
\]

The link between the \( v_{i,j} \)'s and the sufficient statistic of the Potts model is

\[
S(z_k) = \sum_{i \sim j} \mathbb{1}(v_{i,j} = 0).
\]

Recall that \( N_{\text{nei}} \) denotes the number of pairs of neighbor pixels, i.e. the number of vertices.

**Theorem 1.** The values of these \( N_{\text{nei}} \) vertices are fixed only through the values of a subset of them: if we know the values of \( N - 1 \) vertices not forming any loop, we are able to compute the values of the \( N_{\text{nei}} - N + 1 \) other vertices, and thus to compute the value of \( S(z_k) \).

The proof is straightforward. A graph containing \( N \) edges and \( N - 1 \) vertices not forming any loop is a tree (Diestel, 2010): its \( N \) edges are connected. We may fix the component of one of the \( N \) pixels to \( p \) and, since all the pixels are connected, we are able to compute the values of the \( N - 1 \) other pixels, and consequently the values of all of the \( N_{\text{nei}} \) vertices. Since the values of the pixels depend on \( p \) but the values of the vertices do not, the theorem stands.

Therefore, we can produce a random simulation of \( S(z_k) \) through the simulation of \( N - 1 \) unlooping vertices. Let \( N_0 \) denote the number of these \( N - 1 \) unlooping vertices whose value is 0.
When $\beta_k = 0$, $N_0$ follows a Binomial distribution with parameters $N - 1$ and $1/k$. Since

$$P_{\beta_k=0}(S(z_k) = s) = \sum_{i=0}^{N-1} P_{\beta_k=0}(N_0 = i)P_{\beta_k=0}(S(z_k) = s|N_0 = i),$$

the first term being given by the Binomial distribution, we only have to evaluate $P_{0}(S(z_k) = s|N_0 = i)$ in order to estimate $P_{0}(S(z_k) = s)$. The process we chose is the following:

- choose randomly the set of $N - 1$ unlooping vertices,
- select the $N_0$ among them which are set to 0,
- for each of the $N - 1 - N_0$ remaining vertices, simulate its value uniformly on $\{1, \cdots, k-1\}$,
- compute the sufficient statistic $S(z_k)$.

On Figure 3, we have plotted all the couples $(N_0, S(z_k))$ we simulated for $k = 10$. Remark that the conditional probability $P_{\beta_k=0}(S(z_k)|N_0)$ has to be estimated for any value of $N_0$, but only for a small range of $S(z_k)$, the range of values the Gibbs sampler focuses on after convergence. In our exemple, for $K = 10$, this range was $[14947, 15245]$. We decided to estimate the probability $P_{0}(S(z_k) = s|N_0 = i)$ by the empirical frequency computed from the simulated sample.

![Figure 3: Values of $S(z_k)$ depending on $N_0$.](image)

Now we are able to estimate the completed likelihood $f(y, z_k|\theta_k)$, we decide to run the maximization process on the sample issued from the Gibbs sampler after convergence.
5 Numerical illustration on a satellite image of the Lake of Menteith

The values obtained for the ICL criterion are given in Table 1. A number of 10 components is finally chosen. Figure 4 gives the values taken by the Potts model parameter, $\beta_k$, during Gibbs sampling when $k = 10$. This plot highlights the useless of the scheme of Murray et al. (2006) for generating the Potts parameter: it converges very quickly and mixes quite well. Finally, Figure 5 is the plot of the field $z_k^*$ maximizing the complete likelihood. Remark that, contrary to a classical mixture model, here we cannot compute the posterior mode for each pixel because of the spatial dependence.

We also applied the mean-field technique introduced by Forbes and Peyrard (2003). A number of 11 components is chosen. The EM algorithm used by this method leads to very unstable results. For each number of component, we used ten trials. We cannot be sure that the eleven-component model is really the optimal one.

Table 1: ICL criterion

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICL</td>
<td>88286</td>
<td>79564</td>
<td>76647</td>
<td>66874</td>
<td>64174</td>
<td>61620</td>
<td>59874</td>
<td>59699</td>
<td>59580</td>
<td>58401</td>
<td>58515</td>
</tr>
</tbody>
</table>

Figure 4: Values of $\beta_k$. 
References


