A review on statistical inference methods for discrete Markov random fields

Julien Stoehr

▶ **To cite this version:**

Julien Stoehr. A review on statistical inference methods for discrete Markov random fields. 2017. hal-01462078v1

HAL Id: hal-01462078

https://hal.science/hal-01462078v1

Preprint submitted on 8 Feb 2017 (v1), last revised 11 Apr 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A review on statistical inference methods for discrete Markov random fields

Julien Stoehr[1]

[1]School of Mathematical Sciences & Insight Centre for Data Analytics, University College Dublin, Ireland

February 8, 2017

Abstract

Developing satisfactory methodology for the analysis of Markov random field is a very challenging task. Indeed, due to the Markovian dependence structure, the normalizing constant of the fields cannot be computed using standard analytical or numerical methods. This forms a central issue for any statistical approach as the likelihood is an integral part of the procedure. Furthermore, such unobserved fields cannot be integrated out and the likelihood evaluation becomes a doubly intractable problem. This report gives an overview of some of the methods used in the literature to analyse such observed or unobserved random fields.

Keywords: statistics; Markov random fields; parameter estimation; model selection.

1 Introduction

The problem of developing satisfactory methodology for the analysis of spatial data has been of a constant interest for more than half a century now. Constructing a joint probability distribution to describe the global properties of data is somewhat complicated but the difficulty can be bypassed by specifying the local characteristics via conditional probability instead. This proposition has become feasible with the introduction of Markov random fields (or Gibbs distribution) as a family of flexible parametric models for spatial data (*the Hammersley-Clifford*

theorem, Besag, 1974). Markov random fields are spatial processes related to lattice structure, the conditional probability at each nodes of the lattice being dependent only upon its neighbours, that is useful in a wide range of applications. In particular, hidden Markov random fields offer an appropriate representation for practical settings where the true state is unknown. The general framework can be described as an observed data $\mathbf{y}$ which is a noisy or incomplete version of an unobserved discrete latent process $\mathbf{x}$.

Gibbs random fields originally come from physics (*see for example*, Lanford and Ruelle, 1969) but have been useful in many other modelling areas, surged by the development in the statistical community since the 1970's. Indeed, they have appeared as convenient statistical model to analyse different types of spatially correlated data. Notable examples are the autologistic model (Besag, 1974) and its extension the Potts model. Shaped by the development of Geman and Geman (1984) and Besag (1986), – see for example Alfò et al. (2008) and Moores et al. (2014) who performed image segmentation with the help of this modelling – and also in other applications including disease mapping (*e.g.*, Green and Richardson, 2002) and genetic analysis (*e.g.*, François et al., 2006, Friel et al., 2009) to name a few. The exponential random graph model or p^* model (Wasserman and Pattison, 1996) is another prominent example (Frank and Strauss, 1986) and arguably the most popular statistical model for social network analysis (*e.g.*, Robins et al., 2007).

Interests in these models is not so much about Markov laws that may govern data but rather the flexible and stabilizing properties they offer in modelling. Despite its popularity, the Gibbs distribution suffers from a considerable computational curse since its normalizing constant is of combinatorial complexity and generally can not be evaluated with standard analytical or numerical methods. This forms a central issue in statistical analysis as the computation of the likelihood is an integral part of the procedure for both parameter inference and model selection. Remark the exception of small latices on which we can apply the recursive algorithm of Reeves and Pettitt (2004), Friel and Rue (2007) and obtain an exact computation of the normalizing constant. However, the complexity in time of the aforementioned algorithm grows exponentially and is thus helpless on large lattices.. Many deterministic or stochastic approximations have been proposed for circumventing this difficulty and developing methods that are computationally efficient and accurate is still an area of active research.

The present survey paper cares about the problem of carrying out statistical inference (mostly in a Bayesian framework) for Markov random fields. When dealing with hidden random fields, the focus is solely on hidden data represented by discrete models such as the Ising or the Potts models. Both are widely used examples and representative of the general level of difficulty. Aims may be to infer on parameters of the model or on the latent state $\mathbf{x}$. The paper is organised as follows: it begins by introducing the existence of Markov random fields with some specific examples (Section 2). The difficulties inherent to the analysis of such a stochastic model are espe-

cially pointed out in the latter. As befits a survey paper, Section 5 and Section 6 are dedicated to a state of the art concerning parameter inference and Section 9 is related to model selection.

2 Markov random field and Gibbs distribution

2.1 Gibbs-Markov equivalence

A discrete random field $\mathbf{X}$ is a collection of random variables X_i indexed by a finite set $\mathscr{S} = \{1,\dots,n\}$, whose elements are called sites, and taking values in a finite state space $\mathscr{X}_i$. For a given subset $A \subset \mathscr{S}$, $\mathbf{X}_A$ and $\mathbf{x}_A$ respectively define the random process on A, *i.e.*, $\{X_i, i \in A\}$, and a realisation of $\mathbf{X}_A$. Denotes $\mathscr{S} \setminus A = -A$ the complement of A in $\mathscr{S}$.

When modelling local interactions, the sites are lying on an undirected graph $\mathscr{G}$ which induces a topology on $\mathscr{S}$: by definition, sites i and j are adjacent or neighbour if and only if i and j are linked by an edge in $\mathscr{G}$. A random field $\mathbf{X}$ is a Markov random field with respect to $\mathscr{G}$, if for all configuration $\mathbf{x}$ and for all sites i

$$\mathbf{P}\left(X_i = x_i \mid \mathbf{X}_{-i} = \mathbf{x}_{-i}\right) = \mathbf{P}\left(X_i = x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \mathbf{x}_{\mathscr{N}(i)}\right), \tag{2.1}$$

where $\mathscr{N}(i)$ denotes the set of all the adjacent sites to i in $\mathscr{G}$. The property (2.1) is a Markov property – the random variable at a site i is conditionally independent of all other sites in $\mathscr{S}$, given its neighbours values – that extends the notion of Markov chains to spatial data. It is worth noting that any random field is a Markov random field with respect to the trivial topology, that is the cliques of $\mathscr{G}$ are either the empty set or the entire set of sites $\mathscr{S}$. Recall a clique c in an undirected graph $\mathscr{G}$ is any single vertex or a subset of vertices such that every two vertices in c are connected by an edge in $\mathscr{G}$. As an example, when modelling a digital image, the lattice is interpreted as a regular 2D-grid of pixels and the random variables states as shades of grey or colors. Two widely used adjacency structures are the graph $\mathscr{G}_4$ (first order lattice), respectively $\mathscr{G}_8$ (second order lattice), for which the neighbourhood of a site is composed of the four, respectively eight, closest sites on a two-dimensional regular lattice, except on the boundaries of the lattice, see Figure 1.

The difficulty with the Markov formulation is that one defines a set of conditional distributions which does not guarantee the existence of a joint distribution. The Hammersley-Clifford theorem states that if the distribution of a Markov random field with respect to a graph $\mathscr{G}$ is positive for all configuration $\mathbf{x}$ then it admits a Gibbs representation for the same topology (see

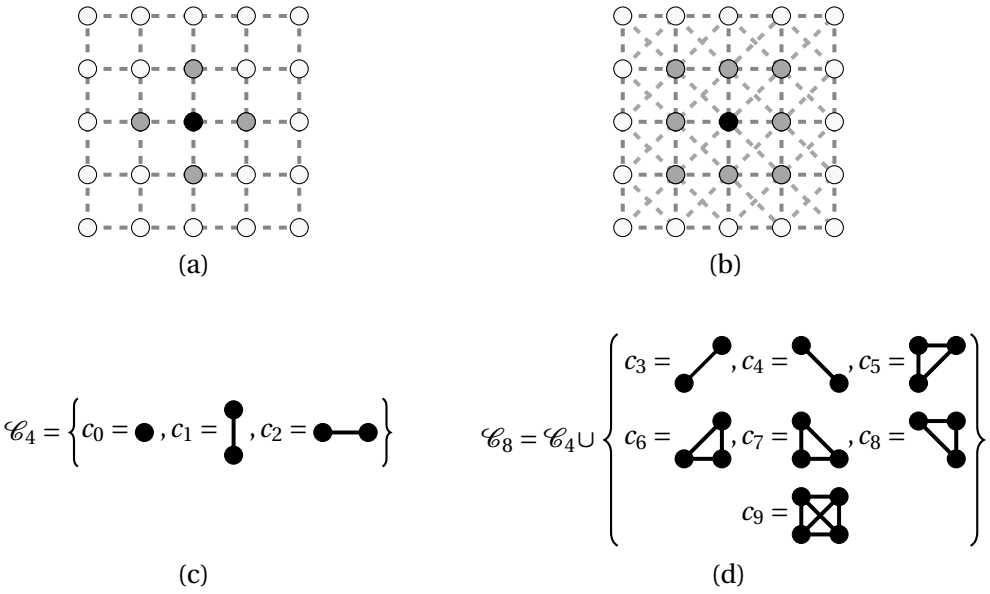

Figure 1: First and second order neighbourhood graphs $\mathscr{G}$ with corresponding cliques. (a) The four closest neighbours graph $\mathscr{G}_4$. neighbours of the vertex in black are represented by vertices in gray. (b) The eight closest neighbours graph $\mathscr{G}_8$. neighbours of the vertex in black are represented by vertices in gray. (c) Cliques of graph $\mathscr{G}_4$. (d) Cliques of graph $\mathscr{G}_8$.

*e.g.*Grimmett (1973), Besag (1974) and for a historical perspective Clifford (1990)), namely a probability measure π on the configuration space $\mathscr{X} = \prod_{i=1}^{n} \mathscr{X}_i$ with the following representation

$$\pi\left(\mathbf{x} \,|\, \psi, \mathscr{G}\right) = \frac{1}{Z\left(\psi, \mathscr{G}\right)} \exp\left\{-H\left(\mathbf{x} \,|\, \psi, \mathscr{G}\right)\right\}, \tag{2.2}$$

where ψ is a vector of parameters, H denotes the energy function or Hamiltonian which can be written as a sum of potential functions over the cliques c of the graph $\mathscr{G}$,

$$H\left(\mathbf{x} \,|\, \psi, \mathscr{G}\right) = \sum_{c} V_c\left(\mathbf{x}_c, \psi\right), \tag{2.3}$$

and $Z\left(\psi, \mathscr{G}\right)$ designates the normalizing constant, called the partition function,

$$Z\left(\psi, \mathscr{G}\right) = \int_{\mathscr{X}} \exp\left\{-H\left(\mathbf{x} \,|\, \psi, \mathscr{G}\right)\right\} \mu(\mathrm{d}\mathbf{x}). \tag{2.4}$$

where μ is the counting measure (discrete case) or the Lebesgue measure (continuous case).

The primary interest of Gibbs distributions comes from statistical physics to describe equilibrium state of a physical systems which consists of a very large number of interacting particles such as ferromagnet ideal gases (Lanford and Ruelle, 1969). Gibbs distribution actually represents disorder system that maximizes the entropy

$$\mathbf{S}(\mathbf{P}) = -\mathbf{E}\{\log \mathbf{P}\} = -\int_{\mathscr{X}} \log \mathbf{P} \mathrm{d}\mathbf{P}$$

over the set of probability distribution $\mathbf{P}$ on configuration space $\mathscr{X}$ with the same expected energy $\mathbf{E}\{H(\mathbf{X} \mid \psi, \mathscr{G})\} = \int_{\mathscr{X}} H(\cdot \mid \psi, \mathscr{G}) \, \mathrm{d}\mathbf{P}$. Ever since, Gibbs random fields have been widely used to analyse different types of spatially correlated data with a wide range of applications, including image analysis (*e.g.,* Hurn et al., 2003, Alfò et al., 2008, Moores et al., 2014), disease mapping (*e.g.,* Green and Richardson, 2002), genetic analysis (François et al., 2006) among others (*e.g.,* Rue and Held, 2005).

Whilst the Gibbs-Markov equivalence provides an explicit form of the joint distribution and thus a global description of the model, this is marred by major difficulties. Conditional probabilities can be easily computed from the likelihood (2.2), but the joint and the marginal distribution are meanwhile unavailable due to the intractable partition function (2.4). For instance in the discrete case, the normalizing constant is a summation over all the possible configurations $\mathbf{x}$ and thus implies a combinatory complexity. For binary variables X_i, the number of possible configurations reaches 2^n.

2.2 Autologistic model and related distributions

The Hammersley-Clifford theorem provides valid probability distributions associated with the random variables $X_1, \ldots, X_n$ by simply specifying local dependency and leads to a class of flexible parametric models for spatial data. In most cases, cliques c_0 of size one (singleton) and cliques $C = \{c_1, c_2, c_3, c_4\}$ of size two (doubleton) are assumed to be satisfactory to model the spatial dependency and potential functions related to larger cliques are set to zero in (2.3). The latter are referred as pairwise Markov random fields. When the full-conditional distribution of each sites belongs to the exponential family, the models deriving from that energy function are the so-called auto-models of Besag (1974). Examples thereafter correspond to Hamiltonian which linearly depends on the parameter $\psi = \{\alpha, \beta\} \in \mathbb{R}^p \times \mathbb{R}^q$, where α scales the potential on sites while β set the strength of interaction between neighbouring pairs, that is,

$$H(\mathbf{x} \mid \alpha, \beta, \mathscr{G}) = V_{c_0}(\mathbf{x}, \alpha) + \sum_{c \in C} V_c(\mathbf{x}, \beta) = -\alpha^T \mathbf{R}(\mathbf{x}) - \beta^T \mathbf{S}(\mathbf{x}).$$

Autologistic model The autologistic model first proposed by Besag (1972) is a pairwise-interaction Markov random field for binary (zero-one) spatial process. The joint distribution is given by

$$\pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \frac{1}{Z\left(\psi, \mathscr{G}\right)} \exp\left\{ \alpha \sum_{i=1}^{n} x_i + \sum_{i \overset{\mathscr{G}}{\sim} j} \beta_{ij} x_i x_j \right\}. \tag{2.5}$$

The full-conditional probability thus writes

$$\pi\left(x_i \,\middle|\, \mathbf{x}_{\mathscr{N}(i)}, \psi, \mathscr{G}\right) = \frac{\exp\left\{\alpha x_i + \sum_{i \overset{\mathscr{G}}{\sim} j} \beta_{ij} x_i x_j\right\}}{1 + \exp\left\{\alpha + \sum_{i \overset{\mathscr{G}}{\sim} j} \beta_{ij} x_j\right\}},$$

and is like a logistic regression where the explanatory variables are the neighbours and themselves observations. The parameter α controls the level of $0-1$ whereas the parameters $\{\beta_{ij}\}$ model the dependency between two neighbouring sites i and j.

One usually prefers to consider variables taking values in $\{-1, 1\}$ instead of $\{0, 1\}$ since it offers a more parsimonious parametrisation and avoids non-invariance issues when one switches states 0 and 1 as mentioned by Pettitt et al. (2003). Note the model stays autologistic but the full-conditional probability turns into

$$\pi\left(x_i \,\middle|\, \mathbf{x}_{\mathscr{N}(i)}, \psi, \mathscr{G}\right) = \frac{\exp\left\{2\alpha x_i + 2\sum_{i \overset{\mathscr{G}}{\sim} j} \beta_{ij} x_i x_j\right\}}{1 + \exp\left\{2\alpha + 2\sum_{i \overset{\mathscr{G}}{\sim} j} \beta_{ij} x_j\right\}}.$$

A well known example is the general Ising model of ferromagnetism (Ising, 1925) that consists of discrete variables representing spins of atoms. The Gibbs distribution (2.5) is referred to as the Boltzmann distribution in statistical physics. The potential on singletons describes local contributions from external fields to the total energy. Spins most likely line up in the same direction of α, that is, in the positive, respectively negative, direction if $\alpha > 0$, respectively $\alpha < 0$. When $\alpha = 0$, there is no external influence. Putting differently α adjusts non-equal abundances of the two state values. The parameters $\{\beta_{ij}\}$ represent the interaction strength between neighbours i and j. When $\beta_{ij} > 0$ the interaction is called ferromagnetic and adjacent spins tend to be aligned, that is neighbouring sites with same sign have higher probability. When $\beta_{ij} < 0$ the interaction is called anti-ferromagnetic and adjacent spins tend to have opposite signs. When $\beta_{ij} = 0$, the spins are non-interacting.

Potts model The Potts model (Potts, 1952) is a pairwise Markov random field that extends the Ising model to K possibles states. The model sets a probability distribution on $\mathbf{x}$ parametrized

by ψ, namely

$$\pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \frac{1}{Z\left(\psi, \mathscr{G}\right)} \exp\left\{ \sum_{i=1}^{n} \sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \sum_{i \overset{\mathscr{G}}{\sim} j} \beta_{ij} \mathbf{1}\{x_i = x_j\} \right\}, \tag{2.6}$$

where $\mathbf{1}\{A\}$ is the indicator function equal to 1 if A is true and 0 otherwise. For instance, as regards the interaction parameter β_{ij}, the indicator function takes the value 1 if the two lattice points i and j take the same value, and 0 otherwise. Note that a potential function can be defined up to an additive constant. To ensure that potential functions on singletons are uniquely determined, one usually imposes the constraint $\sum_{k=0}^{K-1} \alpha_k = 0$.

For $K = 2$, the Potts model is equivalent to the Ising model up to a constant. This is perhaps more transparent by rewriting the Ising model. Consider $\tilde{\mathbf{x}}$ a configuration of the Ising model and assume now $\alpha = \alpha_1 = -\alpha_0$,

(i) for any site i, $\alpha \tilde{x}_i = \alpha_0 \mathbf{1}\{\tilde{x}_i = -1\} + \alpha_1 \mathbf{1}\{\tilde{x}_i = 1\}$,

(ii) for any neighbouring sites i and j, $\tilde{x}_i \tilde{x}_j = 2\mathbf{1}\{\tilde{x}_i = \tilde{x}_j\} - 1$.

The transformation $\tilde{\mathbf{x}} = 2\mathbf{x} - 1$ allows then to conclude. One shall remark here interaction parameters are slightly different between Potts and Ising model. To obtain the same strength of interaction in both model, parameters should satisfy $\beta_{\text{Potts}} = 2\beta_{\text{Ising}}$.

In the literature, one often uses these models in their simplified versions, that is, isotropic ($\beta \in \mathbb{R}$) and without any external field ($\alpha = 0$). For the sake of clarity, I keep the same convention in what follows unless otherwise specified, namely

$$\underline{\text{Ising}}\text{:}\quad \pi\left(\mathbf{x} \,\middle|\, \beta, \mathscr{G}\right) = \frac{1}{Z\left(\beta, \mathscr{G}\right)} \exp\left\{ \beta \sum_{i \overset{\mathscr{G}}{\sim} j} x_i x_j \right\}, \tag{2.7}$$

$$\underline{\text{Potts}}\text{:}\quad \pi\left(\mathbf{x} \,\middle|\, \beta, \mathscr{G}\right) = \frac{1}{Z\left(\beta, \mathscr{G}\right)} \exp\left\{ \beta \sum_{i \overset{\mathscr{G}}{\sim} j} \mathbf{1}\{x_i = x_j\} \right\}. \tag{2.8}$$

2.3 Phase transition

One major peculiarity of Markov random field is a symmetry breaking for large values of parameter β due to a discontinuity of the partition function when the number of sites n tends to

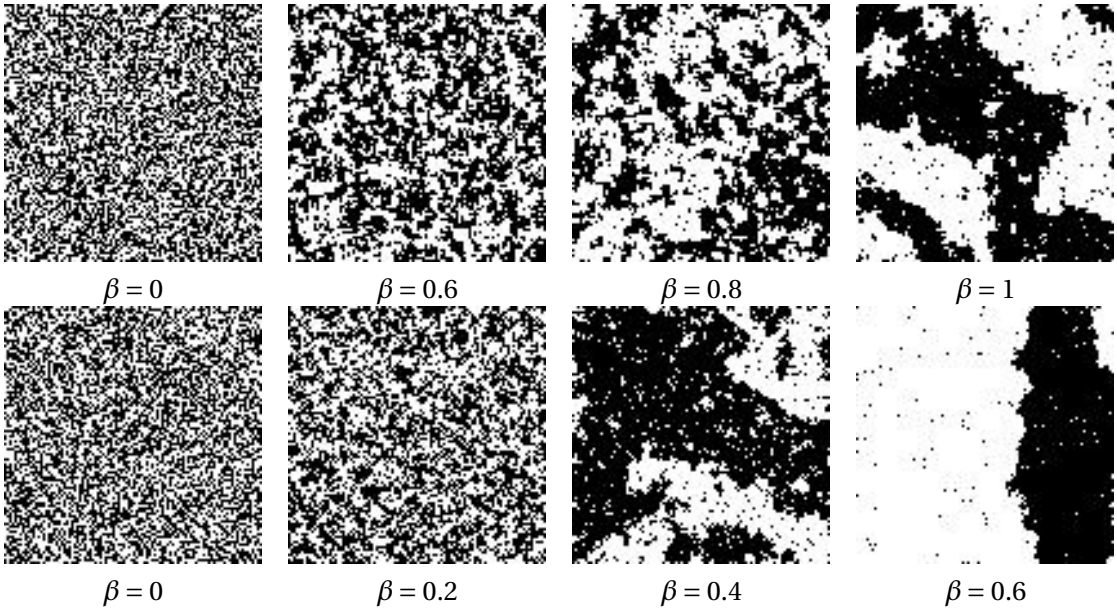

Figure 2: Realization of a 2-states Potts model for various interaction parameter β on a 100×100 lattice with a first-order neighbourhood (first row) or a second-order neighbourhood (second row).

infinity. In physics this is known as phase transition. This transition phenomenon has been widely study in both physics and probability, see for example Georgii (2011) for further details. This part gives particular results for Ising and Potts models on a rectangular lattice.

As already mentioned, the parameter β controls the strength of association between neighbouring sites (see Figure 2). When the parameter β is zero, the random field is a system of independent uniform variables and all configurations are equally distributed. Increasing β favours the variable X_i to be equal to the dominant state among its neighbours and leads to patches of like-valued variables in the graph, such that once β tends to infinity values x_i are all equal. The distribution thus becomes multi-modal. Mention here, this phenomenon vanishes in the presence of an external field (*i.e.*, $\alpha \neq 0$).

In dimension 2, the Ising model is known to have a phase transition at a critical value β_c. When the parameter is above the critical value, $\beta_c < \beta$, one moves gradually to a multi-modal distribution, that is, values x_i are almost all equal for β sufficiently above the critical value. Onsager (1944) obtained an exact value of β_c for a homogeneous Ising model on the first order square lattice, namely

$$\beta_c = \frac{1}{2}\log\left\{1+\sqrt{2}\right\} \approx 0.44.$$

The latter extends to a Potts model with K states on the first order lattice

$$\beta_c = \log\left\{1+\sqrt{K}\right\},$$

see for instance Matveev and Shrock (1996) for specific results to Potts model on the square lattice and Wu (1982) for a broader overview.

The transition is more rapid than the number of neighbours increases. To illustrate this point, Figure 3 gives the average proportion of homogeneous pairs of neighbours, and the corresponding variance, for 2-states Potts model on the first and second order lattices of size 100×100. Indeed, phase transition corresponds to

$$\beta \to \lim_{n\to\infty} \frac{1}{n} \nabla \log Z\left(\beta, \mathscr{G}\right) \text{ is discontinuous at } \beta_c. \tag{2.9}$$

One can show that

$$\nabla \log Z(\beta,\mathscr{G}) = -\mathbf{E}\{\mathbf{S}(\mathbf{X})\} \quad \text{and} \quad \nabla^2 \log Z(\beta,\mathscr{G}) = \mathbf{Var}\{\mathbf{S}(\mathbf{X})\},$$

where $\mathbf{S}(\mathbf{X}) = \sum_{i \overset{\mathscr{G}}{\sim} j} \mathbf{1}\{X_i = X_j\}$ is the number of homogeneous pairs of a Potts random field $\mathbf{X}$, see Section 6.1. Condition (2.9) can thus be written as

$$\lim_{\beta\to\beta_c} \lim_{n\to\infty} \mathbf{Var}\{\mathbf{S}(\mathbf{X})\} = \infty.$$

Mention this is all theoritical asymptotic considerations and the discontinuity does not show itself on finite lattice realizations but the variance becomes increasingly sharper as the size grows.

2.4 Hidden Gibbs random field

The main purpose of this work is to deal with hidden Markov random field, a framework that has encountered a large interest over the past decade. In hidden Markov random fields, the latent process is observed indirectly through another field; this permits the modelling of noise that may happen upon many concrete situations: image analysis, (*e.g.*, Besag, 1986, Stanford and Raftery, 2002, Celeux et al., 2003, Forbes and Peyrard, 2003, Hurn et al., 2003, Alfò et al.,

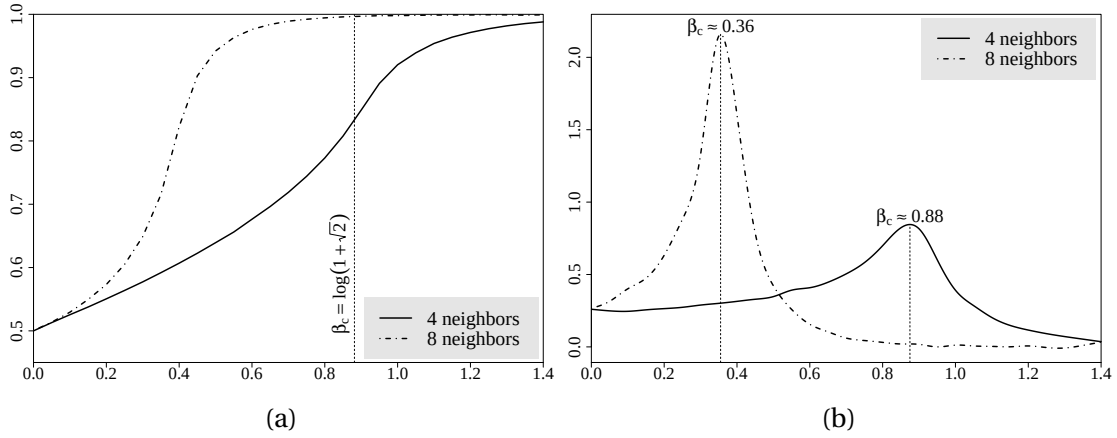

Figure 3: Phase transition for a 2-states Potts model with respect to the first order and second order 100×100 regular square lattices. (a) Average proportion of homogeneous pairs of neighbours. (b) Variance of the number of homogeneous pairs of neighbours.

2008, Friel et al., 2009, Moores et al., 2014), disease mapping (*e.g.,* Green and Richardson, 2002), genetic analysis (François et al., 2006). The aim is to infer some properties of a latent state $\mathbf{x}$ given an observation $\mathbf{y}$. The present part gives a description, in all generality, of the hidden Markov model framework that encompasses the particular cases of hidden Ising or Potts model considered throughout this dissertation.

The unobserved data is modelled as a discrete Markov random field $\mathbf{X}$ associated to an energy function H, as defined in (2.2), parametrized by ψ with state space $\mathscr{X} = \{0, \ldots, K-1\}^n$. Given the realization $\mathbf{x}$ of the latent, the observation $\mathbf{y}$ is a family of random variables indexed by the set of sites $\mathscr{S}$, and taking values in a set $\mathscr{Y}$, *i.e.,* $\mathbf{y} = \left(y_i ; i \in \mathscr{S}\right)$, and are commonly assumed as independent draws that form a noisy version of the hidden field. Consequently, we set the conditional distribution of $\mathbf{Y}$ knowing $\mathbf{X} = \mathbf{x}$, also called emission distribution, as the product

$$\pi\left(\mathbf{y} \,|\, \mathbf{x}, \phi\right) = \prod_{i \in \mathscr{S}} \pi\left(y_i \,|\, x_i, \phi\right),$$

where $\pi(y_i \,|\, x_i, \phi)$ is the marginal noise distribution parametrized by ϕ, that is given for any site i. Those marginal distributions are for instance discrete distributions (Everitt, 2012), Gaussian (*e.g.,* Besag et al., 1991, Qian and Titterington, 1991, Celeux et al., 2003, Forbes and Peyrard, 2003, Friel et al., 2009, Cucala and Marin, 2013) or Poisson distributions (*e.g.,* Besag et al., 1991). Model of noise that takes into account information of the nearest neighbours have also been explored (Besag, 1986).

Assuming that all the marginal distributions $\pi(y_i \mid x_i, \phi)$ are positive, one may write

$$\pi\left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi\right) = \exp\left\{ \sum_{i \in \mathscr{S}} \log \pi\left(y_i \,\middle|\, x_i, \phi\right) \right\},$$

and thus the joint distribution of $(\mathbf{X}, \mathbf{Y})$, also called the complete likelihood, writes as

$$\begin{aligned} \pi\left(\mathbf{x}, \mathbf{y} \,\middle|\, \phi, \psi, \mathscr{G}\right) &= \pi\left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi\right) \pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) \\ &= \frac{1}{Z\left(\psi, \mathscr{G}\right)} \exp\left\{ -H\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) + \sum_{i \in \mathscr{S}} \log \pi\left(y_i \,\middle|\, x_i, \phi\right) \right\}. \end{aligned}$$

The latter equality demonstrates the conditional field $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ is a Markov random field whose energy function satisfies

$$H\left(\mathbf{x} \,\middle|\, \mathbf{y}, \phi, \psi, \mathscr{G}\right) = H\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) - \sum_{i \in \mathscr{S}} \log \pi\left(y_i \,\middle|\, x_i, \phi\right). \tag{2.10}$$

Then, the noise can be interpreted as a non homogeneous external potential on singleton which is a bond to the unobserved data.

3 How to simulate a Markov random field

Sampling from a Gibbs distribution can be a daunting task due to the correlation structure on a high dimensional space, and standard Monte Carlo methods are impracticable except for very specific cases. In the Bayesian paradigm, Markov chain Monte Carlo (MCMC) methods have played a dominant role in dealing with such problems, the idea being to generate a Markov chain whose stationary distribution is the distribution of interest. This section is a reminder of well known algorithms that I make use of throughout numerical parts of this work.

3.1 Gibbs sampler

The Gibbs sampler is a highly popular MCMC algorithm in Bayesian analysis starting with the influential development of Geman and Geman (1984). It can be seen as a component-wise Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970) where variables are updated one at a time and for which proposal distributions are the full conditionals themselves.

It is particularly well suited to Markov random field since the intractable joint distribution is fully determined by the conditional distributions which are easy to compute. Algorithm 1 gives the corresponding algorithmic representation for a joint distribution $\pi(\mathbf{X} \mid \psi, \mathscr{G})$ with a known parameter ψ.

Algorithm 1: Gibbs sampler

Input: a parameter ψ, a number of iterations T
Output: a sample $\mathbf{x}$ from the joint distribution $\pi(\cdot \mid \psi, \mathscr{G})$

Initialization: draw an arbitrary configuration $\mathbf{x}^{(0)} = \left\{x_1^{(0)}, \ldots, x_n^{(0)}\right\}$;
for $t \leftarrow 1$ **to** T **do**
    for $i \leftarrow 1$ **to** n **do**
        draw $x_i^{(t)}$ from the full conditional $\pi\left(X_i^{(t)} \,\middle|\, \mathbf{x}_{\mathscr{N}(i)}^{(t-1)}\right)$;
    end
end
return the configuration $\mathbf{x}^{(T)}$

Geman and Geman (1984, *Theorem A*) have shown the convergence to the target distribution $\pi(\cdot \mid \psi, \mathscr{G})$ regardless of the initial configuration $\mathbf{x}^{(0)}$. The algorithm obviously maintains the target distribution. Says $\mathbf{X}$ has distribution $\pi(\cdot \mid \psi, \mathscr{G})$, at the t-th iteration components of $\mathbf{x}^{(t-1)}$ are replaced by one sampled from the corresponding full conditional distribution induced by $\pi(\cdot \mid \psi, \mathscr{G})$ such that for each of the n steps $\pi(\mathbf{X} \mid \psi, \mathscr{G})$ is stationary. In other words, if $\mathbf{x}$ and $\tilde{\mathbf{x}}$ differ at most from one component i, that is $\mathbf{x}_{-i} = \tilde{\mathbf{x}}_{-i}$, then

$$\sum_{x_i} \pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) \pi\left(\tilde{x}_i \,\middle|\, \mathbf{x}_{-i}, \psi, \mathscr{G}\right) = \pi\left(\tilde{x}_i \,\middle|\, \mathbf{x}_{-i}, \psi, \mathscr{G}\right) \pi\left(\mathbf{x}_{-i} \,\middle|\, \psi, \mathscr{G}\right) = \pi\left(\tilde{\mathbf{x}} \,\middle|\, \psi, \mathscr{G}\right).$$

Under the irreducibility assumption, the chain converges to $\pi(\mathbf{X} \mid \psi, \mathscr{G})$. Note the order in which the components are updated in Algorithm 1 does not make much difference as long as every site is visited. Hence it can be deterministically or randomly modified, especially to avoid possible bottlenecks when visiting the configuration space. A synchronous version is nonetheless unavailable since updating the sites merely at the end of cycle t would lead to incorrect limiting distribution.

We should mention here that Gibbs sampler faces some well known difficulties when it is applied to the Ising or Potts model. The Markov chain mixes slowly, namely long range interactions require many iterations to be taken into account, such that switching the color of a large homogeneous area is of low probability even if the distribution of the colors is exchangeable. This peculiarity is even worse when the parameter β is above the critical value of the phase transition, the Gibbs distribution being severely multi-modal (each mode corresponding to a single color

configuration). Liu (1996) proposed a modification of the Gibbs sampler that overcome these drawbacks with a faster rate of convergence. Note also that in the context of Gaussian Markov random field some efficient algorithm have been proposed like the fast sampling procedure of Rue (2001).

3.2 Auxiliary variables and Swendsen-Wang algorithm

An appealing alternative to bypass slow mixing issues of the Gibbs sampler is the Swendsen-Wang algorithm (Swendsen and Wang, 1987) originally designed to speed up simulation of Potts model close to the phase transition. This algorithm makes a use of auxiliary variables in order to incorporate simultaneous updates of large homogeneous regions (*e.g.,* Besag and Green, 1993). This part describes the procedure for the Potts model with homogeneous external field (2.6).

Denote $\mathbf{x}$ the current configuration of a Markov random field $\mathbf{X}$. Auxiliary random variables aim at decoupling the complex dependence structure between the component of $\mathbf{x}$. Hence we set binary (0-1) conditionally independent auxiliary variables U_{ij} which satisfy

$$\mathbf{P}\left(U_{ij}=1 \,\middle|\, \mathbf{x}\right) = \begin{cases} 1-\exp\left(\beta_{ij}\mathbf{1}\{x_i=x_j\}\right)=p_{ij} & \text{if } i \overset{\mathscr{G}}{\sim} j, \\ 0 & \text{otherwise} \end{cases}$$

with $\beta_{ij} \geq 0$ so that p_{ij} takes value between 0 and 1. The latter then represents the probability to keep an egde between neighbouring sites in $\mathscr{G}$.

The Swendsen-Wang algorithm iterates two steps : a clustering step and a swapping step, see Algorithm 2. Given the configuration $\mathbf{x}$, auxiliary variables yield a partition of sites into single-valued clusters or connected components. Consider the subgraph $\Gamma(\mathscr{G},\mathbf{x})$ of the graph $\mathscr{G}$ induced by U_{ij} on $\mathbf{x}$, namely the undirected graph made of edges of $\mathscr{G}$ for which $U_{ij}=1$, see Figure 4, two sites belong to the same cluster if and only if there is a path between them in $\Gamma(\mathscr{G},\mathbf{x})$. Then each cluster $\mathscr{C}$ is assigned to a new state k with probability

$$\mathbf{P}\left(\mathbf{X}_{\mathscr{C}}=k\right) \propto \exp\left\{\sum_{i\in\mathscr{C}} \alpha_k\right\},$$

where α_k is the component of α associated to the state k. We shall note that for the special but important case where $\alpha=0$, new possible states are equally likely. Also for large values of β, the algorithm manages to switch colors of wide areas, achieving a better cover of the configuration space.

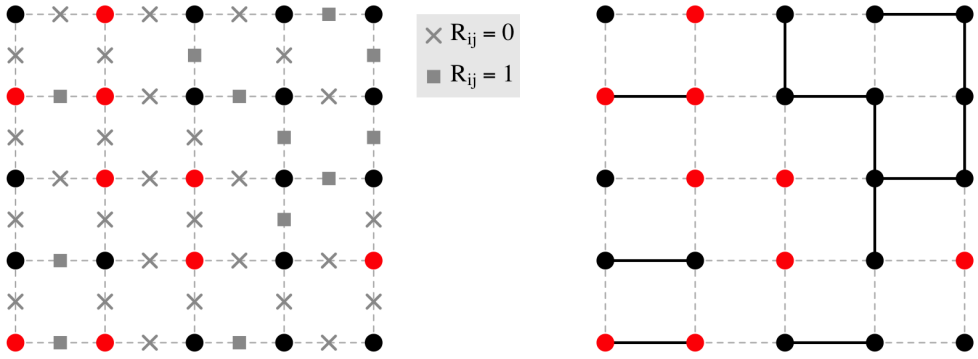

Figure 4: Auxiliary variables and subgraph illustrations for the Swendsen-Wang algorithm. (a) Example of auxiliary variables U_{ij} for a 2-states Potts model configuration on the first order square lattice. (b) Subgraph $\Gamma(\mathscr{G}_4, \mathbf{x})$ of the first order lattice $\mathscr{G}_4$ induced by the auxiliary variables U_{ij}.

For the original proof of convergence, refer to Swendsen and Wang (1987) and for further discussion see for example Besag and Green (1993). Whilst the ability to change large set of variables in one step seems to be a significant advantage, this can be marred by a slow mixing time, namely exponential in n (Gore and Jerrum, 1999). The mixing time of the algorithm is polynomial in n for Ising or Potts models with respect to the graphs $\mathscr{G}_4$ and $\mathscr{G}_8$ but only for small enough value of β (Cooper and Frieze, 1999). This was proved independently by Huber (2003) who also derive a diagnostic tool for the convergence of the algorithm to its invariant distribution, namely using a coupling from the past procedure.

It is worth mentioning that the algorithm can be extended to other Markov random field or models (*e.g.,* Edwards and Sokal, 1988, Wolff, 1989, Higdon, 1998, Barbu and Zhu, 2005) but is then not necessarily efficient. In particular, it is not well suited for latent process. The bound to the data corresponds to a non-homogeneous external field that slows down the computation since the clustering step does not make a use of the data. A solution that might be effective is the partial decoupling of Higdon (1993, 1998). More recently, Barbu and Zhu (2005) make a move from the data augmentation interpretation to a Metropolis-Hastings perspective in order to generalize the algorithm to arbitrary probabilities on graphs. Up to my knowledge, it is not straightforward to bound the Markov chain of such modifications and mixing properties are still an open question despite good results in numerical experiments.

Another alternative for lattice models to make large moves in the configuration space is the slice sampling (*e.g.,* Higdon, 1998) that includes auxiliary variables to sample full conditional distributions in a Gibbs sampler. The sampler is found to have good theoretical properties (*e.g.,* Roberts and Rosenthal, 1999, and the references therein) but this possibility has not been adopted in the present work. Especially I could have used the clever sampler of Mira et al. (2001)

Algorithm 2: Swendsen-Wang algorithm

Input: a parameter ψ, a number of iterations T
Output: a sample $\mathbf{x}$ from the joint distribution $\pi(\cdot \mid \psi, \mathcal{G})$

Initialization: draw an arbitrary configuration $\mathbf{x}^{(0)} = \left\{x_1^{(0)}, \ldots, x_n^{(0)}\right\}$;
for $t \leftarrow 1$ **to** T **do**
  Clustering step: turn off edges of $\mathcal{G}$ with probability $\exp\left(\beta_{ij}\mathbf{1}\{x_i^{(t)} = x_j^{(t)}\}\right)$;
  `// yields the subgraph` $\Gamma\left(\mathcal{G}, \mathbf{x}^{(t)}\right)$ `induced by the auxiliary variables, see Figure 4`
  Swapping step: assign a new state k to each connected component $\mathscr{C}$ of $\Gamma\left(\mathcal{G}, \mathbf{x}^{(t)}\right)$ with probability $\mathbf{P}\left(\mathbf{X}_{\mathscr{C}}^{(t)} = k\right) \propto \exp\left\{\sum_{i \in \mathscr{C}} \alpha_k\right\}$;
end
return the configuration $\mathbf{x}^{(T)}$

that provides exact simulations of Potts models.

4 Recursive algorithm for discrete Markov random field

To answer the difficulty of computing the normalizing constant, generalised recursions for general factorisable models such as the autologistic models have been proposed by Reeves and Pettitt (2004). This method applies to lattices with a small number of rows, up to about 20 for an Ising model, and is based on an algebraic simplification due to the reduction in dependence arising from the Markov property. It applies to unnormalized likelihoods that can be expressed as a product of factors, each of which is dependent on only a subset of the lattice sites.

Denote $q(\mathbf{x} \mid \psi, \mathcal{G})$ the unnormalized version of a Gibbs distribution $\pi(\mathbf{x} \mid \psi, \mathcal{G})$ whose state space is $\mathscr{X} = \{0, \ldots, K-1\}^n$. We can write $q(\mathbf{x} \mid \psi, \mathcal{G})$ as

$$q\left(\mathbf{x} \,\middle|\, \psi, \mathcal{G}\right) = \prod_{i=1}^{n-r} q_i\left(\mathbf{x}_{i:i+r} \,\middle|\, \psi, \mathcal{G}\right),$$

where each factor q_i depends on a subset $\mathbf{x}_{i:r} = \{x_i, \ldots, x_{i+r}\}$ of $\mathbf{x}$, where r is defined to be the *lag* of the model. As a result of this factorisation, the summation for the normalizing constant can be represented as

$$Z\left(\psi, \mathcal{G}\right) = \sum_{\mathbf{x}_{n-r:n}} q_{n-r}\left(\mathbf{x}_{n-r:n} \,\middle|\, \psi, \mathcal{G}\right) \cdots \sum_{\mathbf{x}_{1:1+r}} q_1\left(\mathbf{x}_{1:1+r} \,\middle|\, \psi, \mathcal{G}\right).$$

The latter can be computed much more efficiently than the straightforward summation over the K^n possible lattice realisations using the following steps

$$
\begin{aligned}
Z_1\left(\mathbf{x}_{2:1+r}\right) &= \sum_{x_1} q_1\left(\mathbf{x}_{1:1+r}\right), \\
Z_i\left(\mathbf{x}_{i+1:i+r}\right) &= \sum_{x_i} q_i\left(\mathbf{x}_{i:i+r}\right) Z_{i-1}\left(\mathbf{x}_{i:i+r-1}\right), \text{ for all } i \in \{2,\ldots,n-r\}, \\
Z\left(\psi,\mathscr{G}\right) &= \sum_{\mathbf{x}_{n-r+1:n}} Z_{n-r}\left(\mathbf{x}_{n-r+1:n}\right).
\end{aligned}
$$

The complexity of the troublesome summation is significantly cut down since the forward algorithm solely relies on K^r possible configurations. Note that the algorithm of Reeves and Pettitt (2004) was extended in Friel and Rue (2007) to also allow exact draws from $\pi(\mathbf{x} \mid \psi,\mathscr{G})$ for small enough lattices. The reader can find below an example of implementation for the general Potts model.

Example (Potts model with an external field) Consider a rectangular lattice $h \times w = n$, where h stands for the height and w for the width of the lattice, with a first order neighbourhood system $\mathscr{G}_4$ (see Figure 1.(a)). The model distribution is defined as

$$
\pi(\mathbf{x} \mid \psi,\mathscr{G}_4) = \frac{1}{Z(\psi,\mathscr{G}_4)} \exp\left(\sum_{i=1}^{n} \sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \sum_{i \overset{\mathscr{G}_4}{\sim} j} \beta_{ij} \mathbf{1}\{x_i = x_j\} \right).
$$

The minimum *lag* representation for a Potts lattice with a first order neighbourhood occurs for r given by the smaller of the number of rows or columns in the lattice. Without the loss of generality, assume $h \leq w$ and lattice points are ordered from top to bottom in each column and columns from left to right. It is straightforward to write the unnormalized general Potts distribution as

$$
q\left(\mathbf{x} \,\middle|\, \psi,\mathscr{G}_4\right) = \prod_{i=1}^{n-h} q_i\left(\mathbf{x}_{i:i+h} \,\middle|\, \psi,\mathscr{G}_4\right),
$$

where

- for all lattice point i except the ones on the last row or last column

$$q_i\left(\mathbf{x}_{i:i+h} \mid \psi, \mathscr{G}_4\right) = \exp\Bigg(\sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \beta_0 \mathbf{1}\{x_i = x_{i+1}\} + \beta_1 \mathbf{1}\{x_i = x_{i+h}\} \Bigg). \quad (4.1)$$

- When lattice point i is on the last row x_{i+1} drops out ot (4.1), that is

$$q_i\left(\mathbf{x}_{i:i+h} \mid \psi, \mathscr{G}_4\right) = \exp\left(\sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \beta_1 \mathbf{1}\{x_i = x_{i+h}\} \right). \quad (4.2)$$

- The last factor takes into account all potentials within the last column

$$q_{n-h}\left(\mathbf{x}_{n-h:n} \mid \psi, \mathscr{G}_4\right) = \exp\Bigg(\sum_{i=n-h}^{n} \sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \beta_1 \mathbf{1}\{x_{n-h} = x_n\} + \beta_0 \sum_{i=n-h+1}^{n} \mathbf{1}\{x_i = x_{i+1}\} \Bigg).$$

Identifying the number of rows with the smaller dimension of the lattice, the computation time increases by a factor of K for each additional row, but linearly for additional columns.

One shall remark that for a homogeneous random field, factors (4.1) and (4.2) only depend on the value of the random variables $\mathbf{X}_{i:i+h}$ but not on the actual position of the sites. Hence the number of factors to be computed is $2K^h$ instead of $h(w-1)K^h$. In term of implementation that also means factors can be computed for the different possible configurations once upstream the recursion. Furthermore with a first order neighbourhood, factor at a site merely involves its neighbour below and on its right, thereby reducing the number of possible factor to $K^3 + K^2$.

Algorithm 3 presents the scheme I use in my C++ code which is at the core of numerical experiments presented in Chapter **??** and Chapter **??**. Each configuration $\mathbf{x}_{i+1:i+h}$ corresponds to the unique representation of an integer j belonging to $\{0, \ldots, K^h - 1\}$ in the base-K system, namely

$$j = x_{i+1} + x_{i+2}K + \ldots + x_{i+h}K^{h-1}.$$

As already mentioned, it is enough to calculate factors (4.1) and (4.2) on $\{0, \ldots, K-1\}^3$ and $\{0, \ldots, K-1\}^2$ respectively. Using the previous one-to-one correspondence, the following func-

Algorithm 3: Recursive algorithm

Output: The normalizing constant $Z(\psi, \mathscr{G})$

Compute all the possible factors $q(\cdot)$;
for $j \leftarrow 0$ **to** $K^h - 1$ **do**
  compute $Z(j) \leftarrow \sum_{j=0}^{K-1} q(v_3(k,j))$; `// Corresponds to the computation of` $Z_1(\mathbf{x}_{2:1+r})$
end
for $i \leftarrow 2$ **to** $n-h$ **do**
  save $Z_{\text{old}} \leftarrow \big(Z(1), \ldots, Z(K^h-1)\big)$;
  for $j \leftarrow 0$ **to** $K^h - 1$ **do**
    if *i is not on the last row* **then**
      compute $Z(j) \leftarrow \sum_{k=0}^{K-1} q(v_3(k,j)) Z_{\text{old}}(v(k,j))$;
    else
      compute $Z(j) \leftarrow \sum_{k=0}^{K-1} q(v_2(k,j)) Z_{\text{old}}(v(k,j))$;
    end
  end
end
compute $Z_{\text{norm}} \leftarrow \sum_{j=0}^{K^h-1} q(j) Z(j)$;
return the normalizing constant Z_{norm}

tions determine the value of the sites involved in potentials calculation knowing a given state k and an integer j

$$\begin{array}{rcccc} v_2 & : & \{0,\ldots,K-1\} \times \{0,\ldots,K^h-1\} & \to & \{0,\ldots,K-1\}^2 \\ & & (k,j) & \mapsto & (k, x_{i+h}), \end{array}$$

$$\begin{array}{rcccc} v_3 & : & \{0,\ldots,K-1\} \times \{0,\ldots,K^h-1\} & \to & \{0,\ldots,K-1\}^3 \\ & & (k,j) & \mapsto & (k, x_{i+1}, x_{i+h}), \end{array}$$

Hence, the recursion steps are based on the following factors stored for all (k,j) in $\{0,\ldots,K-1\} \times \{0,\ldots,K^h-1\}$

$$\begin{aligned} q\big(v_2(k,j)\big) &= q\left(k, x_{i+h}\right) = \exp\big(\alpha_k + \beta_1 \mathbf{1}\{x_{i+h} = k\}\big), \\ q\big(v_3(k,j)\big) &= q\left(k, x_{i+1}, x_{i+h}\right) = \exp\big(\alpha_k + \beta_0 \mathbf{1}\{x_{i+1} = k\} + \beta_1 \mathbf{1}\{x_{i+h} = k\}\big). \end{aligned}$$

To handle the last column instead of computing $q_{n-h}(\cdot)$ upstream the recursion, the following quantities are stored for all j in $\{0,\dots,K^h-1\}$

$$q(j) = \exp\left(\sum_{i=n-h+1}^{n} \sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \beta_0 \sum_{i=n-h+1}^{n} \mathbf{1}\{x_i = x_{i+1}\} \right). \tag{4.3}$$

Finally, one shall remark that the transition from $Z_i(\mathbf{x}_{i+1:i+r})$ to $Z_{i-1}(\mathbf{x}_{i:i+r-1})$ is based on the transformation

$$\begin{array}{rccc} \upsilon & : & \{0,\dots,K-1\}\times\{0,\dots,K^h-1\} & \to & \{0,\dots,K^h-1\} \\ & & (k,j) & \mapsto & k+K\left(j \pmod{K^h}\right), \end{array}$$

in Algorithm 3.

It is straightforward to extend this algorithm to hidden Markov random field since as already mention in Section 2.4 the noise corresponds to a non homogeneous potential on singleton and hence the model still writes as a general factorisable model. Algorithm 3 remains the same except for a few details. With the exception of factors (4.3), the potential deriving from the noise is not saved but is added at each step of the recursion, that is the computation of $Z(j)$ turns into

$$\begin{aligned} Z(j) &\leftarrow \sum_{k=0}^{K-1} q\left(\upsilon_3(k,j)\right) \pi\left(y_i \,\middle|\, x_i = k, \phi\right), \text{ or} \\ Z(j) &\leftarrow \sum_{k=0}^{K-1} q(\cdot) Z_{\text{old}}\left(\upsilon(k,j)\right) \pi\left(y_i \,\middle|\, x_i = k, \phi\right). \end{aligned}$$

5 Parameter inference: maximum pseudolikelihood estimator

Parameter estimation in the context of Markov random field is extremely challenging due to the intractable normalizing constant. Much attention has been paid in the literature to this problem arising from maximum likelihood estimation as well as Bayesian inference. The present section presents the solution offered by the pseudolikelihood of Besag (1975) from a maximum likelihood perspective. Its use in a Bayesian framework is discussed in Chapter **??**.

Maximum likelihood estimator

Consider a noisy or incomplete observation, say $\mathbf{y}$, of a hidden Markov random field $\mathbf{x}$. Under the statistical model $\pi(\mathbf{x},\mathbf{y} \mid \theta,\mathscr{G})$, a possible estimate of parameter $\theta = (\psi,\phi)$ is the maximum likelihood estimator. It corresponds to the values of model parameters that maximize the probability of $(\mathbf{x},\mathbf{y})$ for the given statistical model, namely

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \ \pi\left(\mathbf{x},\mathbf{y} \,\middle|\, \theta,\mathscr{G}\right).$$

Equivalently, one can maximize the log-likelihood function. The maximization of the complete likelihood is achieved by maximizing independently the marginal distribution of the hidden process and the conditional distribution of the observation,

$$\hat{\phi}_{\mathrm{MLE}} = \arg\max_{\phi} \ \log\pi\left(\mathbf{y} \,\middle|\, \mathbf{x},\phi\right), \tag{5.1}$$

$$\hat{\psi}_{\mathrm{MLE}} = \arg\max_{\psi} \ \log\pi\left(\mathbf{x} \,\middle|\, \psi,\mathscr{G}\right), \tag{5.2}$$

because $\pi(\mathbf{x},\mathbf{y} \mid \theta,\mathscr{G}) = \pi(\mathbf{y} \mid \mathbf{x},\phi)\pi(\mathbf{x} \mid \psi,\mathscr{G})$. The emission distribution $\pi(\cdot \mid \mathbf{x},\phi)$ has generally some simple form that can at least be evaluated point-wise and the maximization (5.1) is straightforward. On the other hand the optimization problem (5.2) cannot be addressed directly since the gradient has no analytical form and cannot be computed exactly.

Maximum pseudolikelihood estimator

One of the earliest approaches to overcome the intractability of (2.4) is the pseudolikelihood (Besag, 1975) which approximates the joint distribution of $\mathbf{x}$ as the product of full-conditional distributions for each site i,

$$f_{\mathrm{pseudo}}\left(\mathbf{x} \,\middle|\, \psi,\mathscr{G}\right) = \prod_{i=1}^{n} \pi\left(x_i \,\middle|\, \mathbf{x}_{-i},\psi,\mathscr{G}\right) = \prod_{i=1}^{n} \frac{\exp\left\{-\sum_{c|i\in c} V_c\left(\mathbf{x}_c,\psi\right)\right\}}{\sum_{\tilde{x}_i} \exp\left\{-\sum_{c|i\in c} V_c\left(\tilde{\mathbf{x}}_c,\psi\right)\right\}}, \tag{5.3}$$

where the sums $\sum_{c|i\in c}$ and $\sum_{\tilde{x}_i}$ range over the set of cliques containing i and all the possible realization of the random variable X_i respectively. For such a given clique c and a given realization

$\tilde{x}_i, \tilde{\mathbf{x}}_c$ denotes the subgraph that differs from $\mathbf{x}_c$ only at sites i, namely $\tilde{\mathbf{x}}_c = \{\tilde{x}_i\} \cup \{x_j,\ j \in c \setminus \{i\}\}$. The property of Markov random fields ensures that each term in the product only involves nearest neighbours, and so the normalising constant of each full-conditional is straightforward to compute. It is worth noting that pseudolikelihood methods are closely related to the coding method (Besag, 1974) but have a lower computational cost. The maximum pseudolikelihood estimator is computed by maximizing the log-pseudolikelihood

$$\hat{\psi}_{\text{MPLE}} = \arg\max_{\psi} \log f_{\text{pseudo}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right).$$

Similarly to (6.3), one can show that a unique maximum exists which can be estimated with a simple optimization algorithm.

The pseudolikelihood (5.3) is not a genuine probability distribution, except if the random variables X_i are independent. Nevertheless it has been used in preference to Monte Carlo methods since it requires no simulations and provides much faster procedures. Though Geman and Graffigne (1986) demonstrate the consistency of the maximum pseudolikelihood estimator when the lattice size tends to infinity for discrete Markov random field, the result does not imply a good behavior at finite lattice size. Indeed this approximation has been shown to lead to unreliable estimates of ψ especially nearby the phase transition (*e.g.*, Geyer, 1991, Rydén and Titterington, 1998, Friel and Pettitt, 2004, Cucala et al., 2009). Considering it behaves poorly, the much greater expense of Monte Carlo estimators presented in Section 6.1 is justified to supersede the maximum pseudolikehood estimate.

6 Parameter inference: computation of the maximum likelihood

Preferably to maximum pseudolikelihood estimates, many solutions have been explored in the literature to provide approximations of the maximum likelihood estimator. Notable contributions have been given by Monte Carlo techniques even if they may have the drawback of being time consuming (*e.g.*, Younes, 1988, Geyer and Thompson, 1992). An alternative broadly exploited in the context of latent variables is the variational Expectation-Maximization-like algorithms based on an approximation of the Gibbs distribution by product distributions (Celeux et al., 2003). The present section is the occasion to present both solutions, which are used in Chapter **??** and Chapter **??**.

6.1 Monte Carlo maximum likelihood estimator

The use of Monte-Carlo techniques in preference to pseudolikelihood to compute maximum likelihood estimates has been especially highlighted by Geyer and Thompson (1992). Assume Gibbs distributions are of the exponential form, *i.e.*, the Hamiltonian linearly depends on the vector of parameters $\psi = (\psi_1, \dots, \psi_d)$, that is

$$H\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = -\psi^T \mathbf{S}(\mathbf{x}),$$

where $\mathbf{S}(\mathbf{x}) = (s_1(\mathbf{x}), \dots, s_d(\mathbf{x}))$ is a vector of sufficient statistics. Such models have a unique maximum likelihood. Indeed the score function for ψ writes as

$$\nabla \log \pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \mathbf{S}(\mathbf{x}) - \nabla \log Z\left(\psi, \mathscr{G}\right).$$

It is straightforward to show that the partial derivatives of the normalizing constant $Z(\psi, \mathscr{G})$ satisfy

$$\frac{\partial}{\partial \psi_j} \log Z\left(\psi, \mathscr{G}\right) = \frac{\sum_{\mathbf{x} \in \mathscr{X}} s_j(\mathbf{x}) \exp\left\{\psi^T \mathbf{S}(\mathbf{x})\right\}}{\sum_{\mathbf{x} \in \mathscr{X}} \exp\left\{\psi^T \mathbf{S}(\mathbf{x})\right\}} = \mathbf{E}_\psi\left\{s_j(\mathbf{X})\right\}, \tag{6.1}$$

where $\mathbf{E}_\psi(s_j(\mathbf{X}))$ denotes the expected value of $s_j(\mathbf{X})$ with respect to $\pi(\cdot \mid \psi, \mathscr{G})$. Hence the score function can be written as a sum of moments of $s(\mathbf{X})$, namely

$$\nabla \log \pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \mathbf{S}(\mathbf{x}) - \mathbf{E}_\psi\left\{s(\mathbf{X})\right\}. \tag{6.2}$$

Taking the partial derivatives of the previous expression yields similar identities for the Hessian matrix of the log-likelihood for ψ,

$$\nabla^2 \log \pi(\mathbf{x} \mid \psi, \mathscr{G}) = -\mathbf{Cov}_\psi\left\{\mathbf{S}(\mathbf{X})\right\}, \tag{6.3}$$

where $\mathbf{Cov}_\psi\left\{\mathbf{S}(\mathbf{X})\right\}$ denotes the covariance matrix of $\mathbf{S}(\mathbf{X})$ with respect to $\pi(\cdot \mid \psi, \mathscr{G})$. The log-likelihood is thus a concave function and the maximum likelihood estimator $\hat{\psi}_{\text{MLE}}$ is the unique zero of the score function $\nabla \log \pi(\mathbf{x} \mid \psi, \mathscr{G})$, namely

$$\hat{\psi}_{\text{MLE}} = \arg\max_{\psi} \log \pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) \Longleftrightarrow \mathbf{S}(\mathbf{x}) - \mathbf{E}_{\hat{\psi}_{\text{MLE}}}\left\{\mathbf{S}(\mathbf{X})\right\} = 0.$$

Hence a solution to solve problem (5.2) is to resort to stochastic approximations on the basis of equation (6.2) (*e.g.*, Younes, 1988, Descombes et al., 1999). Younes (1988) provides a stochastic gradient algorithm converging under mild conditions. At each iteration the algorithm takes the direction of the estimated gradient with a step size small enough. Another approach to compute the maximum likelihood estimation is to use direct Monte Carlo calculation of the likelihood such as the MCMC algorithm of Geyer and Thompson (1992). The convergence in probability of the latter toward the maximum likelihood estimator is proven for a wide range of models including Markov random fields. Following that work, Descombes et al. (1999) derive also a stochastic algorithm that, as opposed to Younes (1988), takes into account the distance to the maximum likelihood estimator using importance sampling.

6.2 Expectation-Maximization algorithm

A method well suited for estimating parameters in the context of latent variables is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). This iterative procedure has encountered a great success especially in the context of independent mixture model or hidden Markov models. When dealing with Gibbs distributions, the method is subject to the inherent difficulties of the model but several solutions have been proposed in the literature. This section is an opportunity to introduce the solutions that will be particularly useful in Chapter **??**.

The EM algorithm is based on complete-likelihood computation. Consider $\theta = (\psi, \phi)$ with ψ the parameter of the hidden process and ϕ the emission parameter. For the statistical model $\pi(\mathbf{y} \mid \theta)$ (referred to as incomplete likelihood in what follows), the maximum likelihood estimator is defined as

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \pi\left(\mathbf{y} \,\middle|\, \theta\right). \tag{6.4}$$

The EM algorithm addresses problem (6.4) by maximizing at iteration t the expected value of the complete log-likelihood with respect to the conditional distribution of the latent $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ at the current value $\theta^{(t)}$. In other words

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \mathbf{E}\left\{\log \pi\left(\mathbf{X}, \mathbf{y} \,\middle|\, \theta, \mathscr{G}\right) \,\middle|\, \mathbf{Y} = y, \theta^{(t)}\right\} \\
&= \arg\max_{\theta} \sum_{\mathbf{x} \in \mathscr{X}} \pi\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta, \mathscr{G}\right) \log \pi\left(\mathbf{x}, \mathbf{y} \,\middle|\, \theta, \mathscr{G}\right) \\
&:= \arg\max_{\theta} Q\left(\theta \,\middle|\, \theta^{(t)}\right).
\end{aligned} \tag{6.5}$$

Proposition 7. *The log-likelihood* $\log \pi\left(\mathbf{y} \mid \theta^{(t)}\right)$ *increases with* t.

Proof. The result relies on a decomposition of the incomplete log-likelihood that takes into account the latent variables. Given a current value $\theta^{(t)}$, the Bayes theorem allows to write the log-likelihood for all θ in Θ as

$$\begin{aligned}\log \pi\left(\mathbf{y} \,\middle|\, \theta, \mathscr{G}\right) &= \log \pi\left(\mathbf{y} \,\middle|\, \theta\right) \sum_{\mathbf{x} \in \mathscr{X}} \pi\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta^{(t)}, \mathscr{G}\right) \\ &= \sum_{\mathbf{x} \in \mathscr{X}} \log\left\{\frac{\pi\left(\mathbf{x}, \mathbf{y} \,\middle|\, \theta, \mathscr{G}\right)}{\pi\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta, \mathscr{G}\right)}\right\} \pi\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta^{(t)}, \mathscr{G}\right) \\ &= \mathbf{E}\left[\log\left\{\frac{\pi\left(\mathbf{X}, \mathbf{y} \,\middle|\, \theta, \mathscr{G}\right)}{\pi\left(\mathbf{X} \,\middle|\, \mathbf{y}, \theta, \mathscr{G}\right)}\right\} \,\middle|\, \mathbf{Y} = y, \theta^{(t)}\right].\end{aligned}$$

Hence, it decomposes into

$$\log \pi\left(\mathbf{y} \,\middle|\, \theta\right) = Q\left(\theta \,\middle|\, \theta^{(t)}\right) - R\left(\theta \,\middle|\, \theta^{(t)}\right),$$

where $R(\theta \mid \theta^{(t)}) = \mathbf{E}\left\{\log \pi(\mathbf{X} \mid \mathbf{y}, \theta, \mathscr{G}) \mid \mathbf{Y} = \mathbf{y}, \theta^{(t)}\right\}$ and $Q(\theta \mid \theta^{(t)})$ is defined in (6.5). Using Jensen's inequality, one can show that $R(\cdot \mid \theta^{(t)})$ reaches its maximum for $\theta^{(t)}$: for all θ in Θ,

$$\begin{aligned}R\left(\theta \,\middle|\, \theta^{(t)}\right) - R\left(\theta^{(t)} \,\middle|\, \theta^{(t)}\right) &\leq \log\left(\mathbf{E}\left\{\frac{\pi\left(\mathbf{X} \,\middle|\, \mathbf{y}, \theta, \mathscr{G}\right)}{\pi\left(\mathbf{X} \,\middle|\, \mathbf{y}, \theta^{(t)}, \mathscr{G}\right)} \,\middle|\, \mathbf{Y} = \mathbf{y}, \theta^{(t)}\right\}\right) \\ &\leq \log\left\{\sum_{\mathbf{x} \in \mathscr{X}} \pi\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta, \mathscr{G}\right)\right\} \leq 0.\end{aligned}$$

It follows from the previous inequality and $\theta^{(t+1)} = \arg\max_\theta Q(\theta \mid \theta^{(t)})$ that

$$\log \pi\left(\mathbf{y} \,\middle|\, \theta^{(t+1)}\right) \geq \log \pi\left(\mathbf{y} \,\middle|\, \theta^{(t)}\right). \quad \square$$

Wu (1983) demonstrated the convergence under regularity conditions of the sequence $\left\{\theta^{(t)}\right\}_{t \geq 0}$ of the EM algorithm toward a local maximum of $\pi(\mathbf{y} \mid \theta)$ when $t \to \infty$. However, as often with optimization algorithms, the procedure may be very sensitive to the initial value and may exhibit slow convergence rate especially if the log-likelihood has saddle points or plateaus. In place of the genuine EM algorithm, some stochastic versions have been proposed for circumventing these limitations such as the Stochastic EM (SEM) algorithm (Celeux and Diebolt, 1985). The latter consists in simulating a configuration $\mathbf{x}^{(t+1)}$ from $\pi(\mathbf{x} \mid \mathbf{y}, \theta^{(t)}, \mathscr{G})$ after the E-step of Algorithm 4. In the M-step, the maximization of the conditional expectation is replaced with

$$\begin{aligned}\phi^{(t+1)} &= \arg\max_{\phi} \log \pi\left(\mathbf{y} \,\middle|\, \mathbf{x}^{(t+1)}, \phi\right), \\ \psi^{(t+1)} &= \arg\max_{\psi} \sum_{i \in \mathscr{S}} \log \pi\left(x_i^{(t+1)} \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \mathbf{x}_{\mathscr{N}(i)}^{(t+1)}, \psi, \mathscr{G}\right).\end{aligned}$$

Algorithm 4: Expectation-Maximization algorithm

Input: an observation $\mathbf{y}$, a number of iterations T
Output: an estimate of the complete likelihood maximum $\hat{\theta}_{\text{MLE}}$

Initialization: start from an initial guess $\theta^{(0)}$ for θ ; `// the maximum pseudolikelihood estimator can be used as an initial value for the spatial component of` θ
for $t \leftarrow 1$ **to** T **do**
  E-step: compute $Q\left(\theta \mid \theta^{(t)}\right)$ the expected value of the complete log-likelihood with respect to the conditional distribution of the latent $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ at the current value $\theta^{(t)}$ as a function of θ;
  M-step: find $\theta^{(t+1)}$ that maximizes $Q\left(\cdot \mid \theta^{(t)}\right)$, *i.e.*, $\theta^{(t+1)} = \arg\max_\theta Q\left(\theta \mid \theta^{(t)}\right)$;
end
return $\theta^{(T)}$

The EM scheme cannot be applied directly to hidden Markov random fields due to the difficulties inherent to the model. The algorithm yields analytically intractable updates. The function Q can be written as

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= \mathbf{E}\left\{\log \pi\left(\mathbf{X}, \mathbf{y} \mid \theta, \mathscr{G}\right) \mid \mathbf{Y} = \mathbf{y}, \theta^{(t)}\right\} \\ &= \underbrace{\mathbf{E}\left\{\log \pi\left(\mathbf{y} \mid \mathbf{X}, \phi\right) \mid \mathbf{Y} = \mathbf{y}, \theta^{(t)}\right\}}_{=Q_1\left(\phi \mid \theta^{(t)}\right)} + \underbrace{\mathbf{E}\left\{\log \pi\left(\mathbf{X} \mid \psi, \mathscr{G}\right) \mid \mathbf{Y} = \mathbf{y}, \theta^{(t)}\right\}}_{=Q_2\left(\psi \mid \theta^{(t)}\right)}. \end{aligned}$$

The first term of the right hand side only depends on the emission parameter whereas the second one solely involves the Gibbs parameter. Both terms can be further developed as

$$\begin{aligned} Q_1\left(\phi \mid \theta^{(t)}\right) &= \mathbf{E}\left\{ \sum_{i \in \mathscr{S}} \log \pi\left(y_i \mid X_i, \phi\right) \,\middle|\, \mathbf{Y} = \mathbf{y}, \theta^{(t)} \right\} \\ &= \sum_{i \in \mathscr{S}} \sum_{x_i} \pi\left(x_i \mid \mathbf{y}, \theta^{(t)}, \mathscr{G}\right) \log \pi\left(y_i \mid x_i, \phi\right), \\ Q_2\left(\psi \mid \theta^{(t)}\right) &= \mathbf{E}\left\{ -\log Z\left(\psi, \mathscr{G}\right) - \sum_c V_c(\mathbf{X}_c, \psi) \,\middle|\, \mathbf{Y} = \mathbf{y}, \theta^{(t)} \right\} \\ &= -\log Z\left(\psi, \mathscr{G}\right) - \sum_c \sum_{\mathbf{x}_c} \pi\left(\mathbf{x}_c \mid \mathbf{y}, \theta^{(t)}, \mathscr{G}\right) V_c\left(\mathbf{x}_c, \psi\right). \end{aligned} \tag{7.1}$$

The evaluation of Q presents two major difficulties. Neither the partition function $Z(\psi, \mathscr{G})$ arising in Q_2 nor the conditional probabilities $\pi(x_i \mid \mathbf{y}, \theta^{(t)}, \mathscr{G})$ and $\pi(\mathbf{x}_c \mid \mathbf{y}, \theta^{(t)}, \mathscr{G})$ in Q_1 and Q_2 respectively can be easily computed. Many stochastic or deterministic schemes have been proposed and an exhaustive state of art could not be presented here. We focus below on variational

EM-like algorithms that will be used in Chapter **??** for approximating model choice criterion. I could also have mentioned attempts such as the Gibbsian-EM (Chalmond, 1989), the Monte-Carlo EM (Wei and Tanner, 1990) or the Restoration-Maximization algorithm (Qian and Titterington, 1991).

Variational EM algorithm

Variational methods refer to a class of deterministic approaches. They consist in introducing a variational function as an approximation to the likelihood in order to solve a simplified version of the optimization problem. In practice, this relaxation of the original issue has shown good performances for approximating the maximum likelihood estimate (Celeux et al., 2003), as well as for Bayesian inference on hidden Potts model (McGrory et al., 2009).

When dealing with Markov random fields, the mean-field EM is the most popular version of variational EM (VEM) algorithms. The basis is to replace the complex Gibbs distribution with a simple tractable model taken from a family of independent distributions. The principle is to consider the E-step as a functional optimization problem over a set $\mathscr{D}$ of probability distributions on the latent space (*e.g.*, Neal and Hinton, 1998). Similarly to the previous decomposition of the incomplete log-likelihood, for any probability distribution $\mathbf{P}$ in $\mathscr{D}$, one can write

$$\log \pi\left(\mathbf{y} \,\middle|\, \theta\right) = \underbrace{\sum_{\mathbf{x}\in\mathscr{X}} \log\left\{\frac{\pi\left(\mathbf{x},\mathbf{y} \,\middle|\, \theta,\mathscr{G}\right)}{\mathbf{P}(\mathbf{x})}\right\}\mathbf{P}(\mathbf{x})}_{=F(\mathbf{P},\theta)} + \underbrace{\sum_{\mathbf{x}\in\mathscr{X}} \log\left\{\frac{\mathbf{P}(\mathbf{x})}{\pi\left(\mathbf{x} \,\middle|\, \mathbf{y},\theta,\mathscr{G}\right)}\right\}\mathbf{P}(\mathbf{x})}_{=\mathrm{KL}\left(\mathbf{P},\, \pi(\cdot|\mathbf{y},\theta,\mathscr{G})\right)}. \tag{7.2}$$

The last KL term denotes the Kullback-Leibler divergence between a given probability distribution $\mathbf{P}$ and the Gibbs distribution $\pi(\cdot \,|\, \mathbf{y},\theta,\mathscr{G})$. The Kullback-Leibler divergence is a measure of the information lost when one approximates $\pi(\cdot \,|\, \mathbf{y},\theta,\mathscr{G})$ with $\mathbf{P}$. Although it is not a true metric, it has the non-negative property with divergence zero if and only if distributions are equal almost everywhere. The function F introduced in (7.2) is then a lower bound for the log-likelihood. The aim of the variational approach is to maximize the function F instead of the function Q by choosing a distribution $\mathbf{P}$ easy to compute and close enough to $\pi(\cdot \,|\, \mathbf{y},\theta,\mathscr{G})$. This shift in the formulation leads to an alternating optimization procedure which can be described as follows: given a current value $(\mathbf{P}^{(t)},\theta^{(t)})$ in $\mathscr{D}\times\Theta$, updates with

$$\mathbf{P}^{(t+1)} = \operatorname*{arg\,max}_{\mathbf{P}\in\mathscr{D}} F\left(\mathbf{P},\theta^{(t)}\right) = \operatorname*{arg\,min}_{\mathbf{P}\in\mathscr{D}} \mathrm{KL}\left(\mathbf{P},\, \pi\left(\cdot \,\middle|\, \mathbf{y},\theta^{(t)},\mathscr{G}\right)\right), \tag{7.3}$$

$$\theta^{(t+1)} = \operatorname*{arg\,max}_{\theta} F\left(\mathbf{P}^{(t+1)},\theta\right) = \operatorname*{arg\,max}_{\theta} \sum_{\mathbf{x}\in\mathscr{X}} \mathbf{P}^{(t+1)}(\mathbf{x}) \log \pi\left(\mathbf{x},\mathbf{y} \,\middle|\, \theta,\mathscr{G}\right). \tag{7.4}$$

The minimization of the Kullback-Leibler divergence over the whole set of probability distributions on $\mathscr{X}$ has an explicit solution which is the conditional distribution $\pi(\cdot \mid \mathbf{y}, \theta, \mathscr{G})$. Then the maximization over Θ corresponds to the maximization of Q and we recover the standard EM scheme. The proposal of VEM to make the E-step tractable is to solve (7.3) over a restricted set $\tilde{\mathscr{D}}$ of probability distributions: the class of independent probability distributions $\mathbf{P}$ that factorize on sites, namely for all $\mathbf{x}$ in $\mathscr{X} = \prod_{i \in \mathscr{S}} \mathscr{X}_i$,

$$\mathbf{P}(\mathbf{x}) = \prod_{i \in \mathscr{S}} \mathbf{P}_i(x_i), \text{ where } \mathbf{P}_i \in \mathscr{M}_1^+(\mathscr{X}_i) \text{ and } \mathbf{P} \in \mathscr{M}_1^+(\mathscr{X}).$$

The mean field approximation is the optimal solution in $\tilde{\mathscr{D}}$, in the sense that it is the closest distribution to the Gibbs distribution that factorizes on sites. Despite the introduction of the relaxation, the M-step remains intractable due to the latent Markovian structure. Indeed functions Q_1 and Q_2 of equations (**??**) and (**??**) are replaced by

$$Q_1^{\text{VEM}}\left(\phi \,\middle|\, \mathbf{P}^{(t)}\right) = \sum_{i \in \mathscr{S}} \sum_{x_i} \mathbf{P}^{(t)}(\mathbf{x}) \log \pi\left(y_i \,\middle|\, x_i, \phi\right), \tag{7.5}$$

$$Q_2^{\text{VEM}}\left(\psi \,\middle|\, \mathbf{P}^{(t)}\right) = -\log Z\left(\psi, \mathscr{G}\right) - \sum_c \sum_{\mathbf{x}_c} \mathbf{P}^{(t)}(\mathbf{x}) V_c(\mathbf{x}_c, \psi). \tag{7.6}$$

The update of the emission parameter $\phi^{(t+1)}$, obtained by maximizing Q_1^{VEM} can often be computed analytically. In contrast, the update of Gibbs parameter still presents computational challenges since it requires either an explicit expression of the partition function or an explicit expression of its gradient. Further algorithms have been suggested to answer the question. Generalizing an idea originally introduced by Zhang (1992), Celeux et al. (2003) have designed a class of VEM-like algorithm that uses mean field-like approximations for both $\pi(\cdot \mid \mathbf{y}, \theta, \mathscr{G})$ and $\pi(\cdot \mid \psi, \mathscr{G})$. To put it in simple terms mean field-like approximations refer to distributions for which neighbours of site i are set to constants. Given a configuration $\tilde{\mathbf{x}}$ in $\mathscr{X}$, the Gibbs distribution $\pi(\cdot \mid \psi, \mathscr{G})$ is replaced by

$$\mathbf{P}^{\text{MF-like}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \prod_{i \in \mathscr{S}} \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}_{\mathscr{N}(i)}, \psi, \mathscr{G}\right).$$

The main difference with the pseudolikelihood (5.3) is that neighbours are not random anymore and setting them to constant values leads to a system of independent variables. From this approximation, the EM path is set up with the corresponding joint distribution approximation

$$\mathbf{P}^{\text{MF-like}}\left(\mathbf{x}, \mathbf{y} \,\middle|\, \theta, \mathscr{G}\right) = \prod_{i \in \mathscr{S}} \pi\left(y_i \,\middle|\, x_i, \phi\right) \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}_{\mathscr{N}(i)}, \psi, \mathscr{G}\right).$$

Algorithm 5: Simulated Field algorithm

Input: an observation $\mathbf{y}$, a number of iterations T
Output: an estimate of the complete likelihood maximum $\hat{\theta}_{\text{MLE}}$

Initialization: start from an initial guess $\theta^{(0)} = \left(\psi^{(0)}, \phi^{(0)}\right)$;
for $t \leftarrow 1$ **to** T **do**
  neighbourhood restoration: draw $\tilde{\mathbf{x}}^{(t)}$ from $\pi\left(\cdot \,\middle|\, \mathbf{y}, \psi^{(t-1)}, \mathscr{G}\right)$;
  E-step: compute

$$\widehat{Q_1}(\phi) := \sum_{i \in \mathscr{S}} \sum_{x_i} \pi\left(x_i \,|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)}, y_i, \theta^{(t-1)}, \mathscr{G}\right) \log \pi\left(y_i \,\middle|\, x_i, \phi\right);$$

$$\widehat{Q_2}(\psi) := \sum_{i \in \mathscr{S}} \sum_{x_i} \pi\left(x_i \,|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)}, y_i, \theta^{(t-1)}, \mathscr{G}\right) \log \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)}, \psi, \mathscr{G}\right);$$

  M-step: set $\theta^{(t)} = \left(\psi^{(t)}, \phi^{(t)}\right)$ where

$$\phi^{(t)} = \arg\max_{\phi} \widehat{Q_1}(\phi) \text{ and } \psi^{(t)} = \arg\max_{\psi} \widehat{Q_2}(\psi);$$

end
return $\theta^{(T)} = \left(\psi^{(T)}, \phi^{(T)}\right)$

Note that this general procedure corresponds to the so-called point-pseudo-likelihood EM algorithm proposed by Qian and Titterington (1991). The updates of ϕ and ψ become fully tractable by replacing $\pi(\cdot \,|\, \mathbf{y}, \theta, \mathscr{G})$ with its approximation that derives from the Bayes formula

$$\begin{aligned}\mathbf{P}^{\text{MF-like}}\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta, \mathscr{G}\right) &= \frac{\pi\left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi\right) \mathbf{P}^{\text{MF-like}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right)}{\mathbf{P}^{\text{MF-like}}\left(\mathbf{y} \,\middle|\, \theta\right)} \\ &= \prod_{i \in \mathscr{S}} \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}_{\mathscr{N}(i)}, y_i, \theta, \mathscr{G}\right).\end{aligned}$$

Then functions Q_1^{VEM} and Q_2^{VEM} of equations (7.5) and (7.6) are replaced with

$$Q_1^{\text{MF-like}}\left(\phi \,\middle|\, \theta^{(t)}\right) = \sum_{i \in \mathscr{S}} \sum_{x_i} \pi\left(x_i \,|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)}, y_i, \theta^{(t)}, \mathscr{G}\right) \log \pi\left(y_i \,\middle|\, x_i, \phi\right),$$

$$Q_2^{\text{MF-like}}\left(\psi \,\middle|\, \theta^{(t)}\right) = \sum_{i \in \mathscr{S}} \sum_{x_i} \pi\left(x_i \,|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)}, y_i, \theta^{(t)}, \mathscr{G}\right) \log \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)}, \psi, \mathscr{G}\right).$$

The flexibility of the approach proposed by Celeux et al. (2003) lies in the choice of the configuration $\tilde{\mathbf{x}}$ that is not necessarily a valid configuration for the model. In this case the Hamiltonian

should be written differently in order to have a proper formulation of the mean-field approximations. It is unnecessary to introduce this notation here and we refer the reader to Celeux et al. (2003) for further details. When the neighbours $\mathbf{X}_{\mathcal{N}(i)}$ are fixed to their mean value, or more precisely $\tilde{\mathbf{x}}$ is set to the mean field estimate of the complete conditional distribution $\pi(\mathbf{x} \mid \mathbf{y}, \theta, \mathcal{G})$, this results in the Mean Field algorithm of Zhang (1992). In practice, Celeux et al. (2003) obtain better performances with their so-called Simulated Field algorithm (see Algorithm 5). In this stochastic version of the EM-like procedure, $\tilde{\mathbf{x}}$ is a realization drawn from the conditional distribution $\pi(\cdot \mid \mathbf{y}, \theta^{(t)}, \mathcal{G})$ for the current value of the parameter $\theta^{(t)}$. The latter is preferred to other methods when dealing with maximum-likelihood estimation for hidden Markov random field.

This extension of VEM algorithms suffers from a lack of theoretical support due to the propagation of the approximation to the Gibbs distribution $\pi(\cdot \mid \psi, \mathcal{G})$. One might advocate in favour of the Monte-Carlo VEM algorithm of Forbes and Fort (2007) for which convergence results are available. However the Simulated Field algorithm provides better results for the estimation of the spatial parameter, as illustrated in Forbes and Fort (2007).

8 Parameter inference: computation of posterior distributions

Bayesian inference faces the same difficulties than maximum likelihood estimation since the computation of the likelihood is integral to the approach. Chapter **??** addresses the problem of computing the posterior parameter distribution when the Markov random field is directly observed. To tackle the obstacle of the intractable normalising constant, recent work have proposed simulation based approaches. This part focuses on the single auxiliary variable method Møller et al. (2006) and the exchange algorithm Murray et al. (2006): a Gibbs-within-Metropolis-Hastings algorithm. Both solutions may suffer from computational difficulties, either a delicate calibration or a high computational cost. Alternatives that are computationally efficient have been proposed by Friel (2012). The author uses composite likelihoods, that generalize the pseudolikelihood introduced in Section 5, within a Bayesian approach. However the approximation produced has a variability significantly lower than the true posterior. Chapter **??** proposes a correction of composite likelihoods that leads to an accurate estimate without being time consuming.

The current overview is devoted to the Bayesian parameter inference when the Markov random field is fully observed. Recent works have tackled the issue of hidden Markov random fields but it would not possible to describe these here. Nevertheless I shall mention only a few like the exchange marginal particle MCMC of Everitt (2012) or the estimation procedure in Cucala and Marin (2013) that are both based on the exchange algorithm of Murray et al. (2006). Though these methods produce accurate results they inherit the drawback of the exchange algorithm.

Finally, I would add in the toolbox solutions that are computationally more efficient like the reduced dependence approximation of Friel et al. (2009) or the variational Bayes scheme of McGrory et al. (2009).

8.0.1 Posterior parameter distribution

From a Bayesian perspective the focus is on the posterior parameter distribution. In Chapter **??**, we are solely interested in making Bayesian inference about unknown parameters knowing an observed discrete Markov random field $\mathbf{x}^{\text{obs}}$. The hidden case involves an additional level of intractability and is not of interest in the present work.

Assume

(i) a prior on the parameter space Ψ, whose density is $\pi(\psi)$ and

(ii) the likelihood of the data $\mathbf{X}$, namely $\pi(\mathbf{x} \mid \psi, \mathscr{G})$.

The posterior parameter distribution is

$$\pi\left(\psi \,\middle|\, \mathbf{x}^{\text{obs}}, \mathscr{G}\right) \propto \pi\left(\mathbf{x}^{\text{obs}} \,\middle|\, \psi, \mathscr{G}\right) \pi(\psi). \tag{8.1}$$

Posterior parameter estimation is called a doubly-intractable problem because both the likelihood function and the normalizing constant of the posterior distribution are intractable.

8.1 The single auxiliary variable method

The single auxiliary variable method (SAVM) introduced by Møller et al. (2006) is an ingenious MCMC algorithm targeting the posterior distribution (8.1). The original motivation arises from the impossibility to implement a standard Metropolis-Hastings for doubly-intractable distributions. Indeed, to draw a sample from the posterior distribution with a Metropolis-Hastings algorithm one needs to evaluate the ratio

$$r\left(\psi' \mid \psi\right) = \frac{\pi\left(\psi' \mid \mathbf{x}, \mathscr{G}\right)}{\pi\left(\psi \mid \mathbf{x}, \mathscr{G}\right)} \frac{\nu\left(\psi \mid \psi'\right)}{\nu\left(\psi' \mid \psi\right)} = \frac{Z\left(\psi, \mathscr{G}\right)}{Z\left(\psi', \mathscr{G}\right)} \frac{\pi(\psi') q\left(\mathbf{x} \mid \psi', \mathscr{G}\right) \nu\left(\psi \mid \psi'\right)}{\pi(\psi) q\left(\mathbf{x} \mid \psi, \mathscr{G}\right) \nu\left(\psi' \mid \psi\right)}, \tag{8.2}$$

where $\nu(\psi \mid \psi')$ is the proposal density for θ and $q(\mathbf{x} \mid \psi, \mathscr{G})$ is the unnormalized Gibbs distribution. A solution, while being time consuming, is to estimate the ratio of the partition functions using path sampling (Gelman and Meng, 1998). Starting from equation (6.1), the path sampling identity writes as

$$\log\left\{\frac{Z\left(\psi_0, \mathscr{G}\right)}{Z\left(\psi_1, \mathscr{G}\right)}\right\} = \int_{\psi_0}^{\psi_1} \mathbf{E}_{\psi}\{\mathbf{S}(\mathbf{X})\}\mathrm{d}\psi.$$

Hence the ratio of the two normalizing constants can be evaluated with numerical integration. For practical purpose, this approach can barely be recommended within a Metropolis-Hastings scheme since each iteration would require to compute a new ratio.

The proposal of Møller et al. (2006) consists in including an auxiliary variable $\mathbf{U}$ which shares the same state space than $\mathbf{X}$ in order to cancel out the cumbersome normalizing constants. Consider the posterior joint distribution for $(\psi, \mathbf{U})$,

$$\pi\left(\psi, \mathbf{u} \,\middle|\, \mathbf{x}, \mathscr{G}\right) \propto \pi\left(\mathbf{u} \,\middle|\, \mathbf{x}, \psi\right) \frac{q\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right)}{Z\left(\psi, \mathscr{G}\right)} \pi(\psi),$$

where $\pi(\cdot \mid \mathbf{x}, \psi)$ is the conditional distribution for the auxiliary variable. The Metropolis-Hastings ratio for the posterior joint distribution can be written as

$$r\left(\psi', \mathbf{u}' \,\middle|\, \psi, \mathbf{u}\right) = \frac{\pi\left(\psi', \mathbf{u}' \,\middle|\, \mathbf{x}, \mathscr{G}\right) \nu\left(\psi, \mathbf{u} \,\middle|\, \psi', \mathbf{u}', \mathbf{x}\right)}{\pi\left(\psi, \mathbf{u} \,\middle|\, \mathbf{x}, \mathscr{G}\right) \nu\left(\psi', \mathbf{u}' \,\middle|\, \psi, \mathbf{u}, \mathbf{x}\right)},$$

where $\nu(\psi', \mathbf{u}' \mid \psi, \mathbf{u}, \mathbf{x})$ denotes the proposal density for $(\psi, \mathbf{U})$. Assuming the proposal takes the form

$$\nu\left(\psi', \mathbf{u}' \,\middle|\, \psi, \mathbf{u}, \mathbf{x}\right) = \nu\left(\psi' \,\middle|\, \psi, \mathbf{x}\right) \nu\left(\mathbf{u}' \,\middle|\, \psi'\right),$$

Møller et al. (2006) suggest to pick out the intractable likelihood as proposal for the auxiliary variable, namely

$$\nu\left(\mathbf{u}' \,\middle|\, \psi'\right) = \frac{1}{Z\left(\psi', \mathscr{G}\right)} q\left(\mathbf{u}' \,\middle|\, \psi', \mathscr{G}\right).$$

Hence the Metropolis-Hastings acceptance becomes fully tractable,

$$r\left(\psi', \mathbf{u}' \mid \psi, \mathbf{u}\right) = \frac{\cancel{Z\left(\psi, \mathscr{G}\right)}}{\cancel{Z\left(\psi', \mathscr{G}\right)}} \frac{q\left(\mathbf{x} \mid \psi', \mathscr{G}\right) \pi\left(\mathbf{u}' \mid \mathbf{x}, \psi'\right) \pi\left(\psi'\right)}{q\left(\mathbf{x} \mid \psi, \mathscr{G}\right) \pi\left(\mathbf{u} \mid \mathbf{x}, \psi\right) \pi\left(\psi\right)} \frac{\nu\left(\psi \mid \psi', \mathbf{x}\right) q\left(\mathbf{u} \mid \psi, \mathscr{G}\right)}{\nu\left(\psi' \mid \psi, \mathbf{x}\right) q\left(\mathbf{u}' \mid \psi', \mathscr{G}\right)} \frac{\cancel{Z\left(\psi', \mathscr{G}\right)}}{\cancel{Z\left(\psi, \mathscr{G}\right)}}.$$

It follows from the above and (8.2) that the SAVM is based on single point importance sampling approximations of the partition functions $Z(\psi, \mathscr{G})$ and $Z(\psi', \mathscr{G})$, namely

$$\hat{Z}\left(\psi, \mathscr{G}\right) = \frac{q\left(\mathbf{u} \mid \psi, \mathscr{G}\right)}{\pi\left(\mathbf{u} \mid \mathbf{x}, \psi\right)} \text{ and } \hat{Z}\left(\psi', \mathscr{G}\right) = \frac{q\left(\mathbf{u}' \mid \psi', \mathscr{G}\right)}{\pi\left(\mathbf{u}' \mid \mathbf{x}, \psi'\right)}.$$

As mentioned by Everitt (2012), any algorithm producing an unbiased estimate of the normalizing constant can thus be used in place of the importance sampling approximation and will lead to a valid procedure.

The idea to apply MCMC methods to situation where the target distribution can be estimated without bias by using an auxiliary variable construction has appeared in the *generalized importance Metropolis-Hasting* of Beaumont (2003) and has then been extented by Andrieu and Roberts (2009). This brings another justification to the SAVM and possible improvement with the use of sequential Monte Carlo samplers (Andrieu et al., 2010).

8.2 The exchange algorithm

Murray et al. (2006) develop this work further with their exchange algorithm. They outline that SAVM can be improved by directly estimating the ratio $\frac{Z(\psi, \mathscr{G})}{Z(\psi', \mathscr{G})}$ instead of using previous single point estimates. The scheme is a Metropolis-within-Gibbs algorithm (see Algorithm 6) that samples from the augmented posterior distribution

$$\pi\left(\psi, \psi', \mathbf{u} \mid \mathbf{x}, \mathscr{G}\right) \propto \pi\left(\psi\right) \nu\left(\psi' \mid \psi\right) \pi\left(\mathbf{x} \mid \psi, \mathscr{G}\right) \pi\left(\mathbf{u} \mid \psi', \mathscr{G}\right).$$

Comparing the acceptance ratio of Algorithm 6 with the Metropolis-Hasting ratio (8.2), we remark that the intractable ratio $\frac{Z(\psi, \mathscr{G})}{Z(\psi', \mathscr{G})}$ is replaced by $\frac{q(\mathbf{u} \mid \psi, \mathscr{G})}{q(\mathbf{u} \mid \psi', \mathscr{G})}$. The latter can be viewed as a single point importance sampling estimate as pointed out by Murray et al. (2006).

In comparison with the exchange algorithm, the SAVM faces a major drawback. Indeed, the method of Møller et al. (2006) depends on the conditional distribution for the auxiliary variable

Algorithm 6: Exchange algorithm

Input: an initial guess $\left(\psi^{(0)}, \psi'^{(0)}, \mathbf{u}^{(0)}\right)$ for ψ, a number of iterations T
Output: a sample drawn from the augmented distribution $\pi\left(\psi, \psi', \mathbf{u} \,\middle|\, \mathbf{x}, \mathscr{G}\right)$

for $t \leftarrow 1$ **to** T **do**
  draw ψ' from $\nu\left(\cdot \,\middle|\, \psi^{(t-1)}\right)$;
  draw $\mathbf{u}$ from $\pi\left(\cdot \,\middle|\, \psi'^{(t)}, \mathscr{G}\right)$;
  compute the Metropolis-Hastings acceptance ratio

$$r\left(\psi' \,\middle|\, \psi^{(t-1)}, \mathbf{u}\right) = \frac{q\left(\mathbf{u} \,\middle|\, \psi^{(t-1)}, \mathscr{G}\right)}{q\left(\mathbf{u} \,\middle|\, \psi', \mathscr{G}\right)} \frac{\pi\left(\psi'\right) q\left(\mathbf{x} \,\middle|\, \psi', \mathscr{G}\right) \nu\left(\psi^{(t-1)} \,\middle|\, \psi'\right)}{\pi\left(\psi^{(t-1)}\right) q\left(\mathbf{x} \,\middle|\, \psi^{(t-1)}, \mathscr{G}\right) \nu\left(\psi' \,\middle|\, \psi^{(t-1)}\right)};$$

  Exchange move: set $\left(\psi^{(t)}, \psi'^{(t)}, \mathbf{u}^{(t)}\right) = \left(\psi', \psi^{(t-1)}, \mathbf{u}\right)$ with probability $\min\left(1, r(\psi' \mid \psi^{(t-1)}, \mathbf{u})\right)$, else set $\left(\psi^{(t)}, \psi'^{(t)}, \mathbf{u}^{(t)}\right) = \left(\psi^{(t-1)}, \psi'^{(t-1)}, \mathbf{u}^{(t-1)}\right)$;
end
return $\left\{\left(\psi^{(t)}, \psi'^{(t)}, \mathbf{u}^{(t)}\right)\right\}_{t=1}^{T}$

U, namely $\pi(\cdot \mid \mathbf{x}, \psi)$, that makes it difficult to calibrate (see for example Cucala et al., 2009). As a suitable choice for the conditional distribution, the authors advocate in favour of the Gibbs distribution taken at a preliminary estimate $\hat{\psi}$, such as the maximum pseudolikehood, that is

$$\pi\left(\mathbf{u} \,\middle|\, \mathbf{x}, \psi\right) = \frac{1}{Z\left(\hat{\psi}, \mathscr{G}\right)} q\left(\mathbf{u} \,\middle|\, \hat{\psi}, \mathscr{G}\right).$$

By plugging in a particular value $\hat{\psi}$, the normalizing constant $Z(\hat{\psi}, \mathscr{G})$ drops out of the acceptance ratio $r(\psi', \mathbf{u}' \mid \psi, \mathbf{u})$. Nevertheless Cucala et al. (2009) stress out that the choice of $\hat{\psi}$ is paramount and may significantly affect the performances of the algorithm. In this sense, the exchange algorithm is more convenient to implement whilst outperforming the SAVM in Murray et al. (2006).

A practical difficulty remains to implement the exchange algorithm. An exact draw $\mathbf{u}$ from the likelihood $\pi(\cdot \mid \psi, \mathscr{G})$ is required. This is generally infeasible when dealing with Markov random fields with the exception of a very few instances. The Ising model is one of these special cases where $\mathbf{u}$ can be drawn exactly using coupling from the past (Propp and Wilson, 1996) but the perfect simulation may be very expensive especially if the parameter is close to the phase transition. Alternatively, one can run enough iterations of a suitable MCMC (such as Gibbs sampler, Swendsen-Wang algorithm) to reach its stationnary distribution $\pi(\cdot \mid \psi, \mathscr{G})$. This approach has shown good performances in practice (*e.g.*, Cucala et al., 2009, Caimo and Friel, 2011, Everitt, 2012). A theoretical justification is presented by Everitt (2012) who notably pointed out that

solely few iterations of the MCMC sampler are necessary.

9 Model selection

Selecting the model that best fits an observation among a collection of Markov random fields is a daunting task. The comparison of stochastic models is usually based on the Bayes factor (Kass and Raftery, 1995) that is intractable due to a high-dimensional integral. The present dissertation is especially interested in selecting the neighbourhood structure and/or the number of components of hidden discrete Markov random fields such as the hidden Potts model. Approximate Bayesian computation introduced in Section 9.2 brings a solution in the Bayesian paradigm which is explored in Chapter **??**. But it suffers from slow execution. The Bayesian Information Criterion (BIC), which is a simple function of the intractable likelihood at its maximum, is introduced in Section 9.3 and discussed further in Chapter **??**.

9.1 Bayesian model choice

The peculiarity of the Bayesian approach to model selection is to consider the model itself as an unknown parameter of interest. Assume we are given a set $\mathscr{M} = \{m : 1, \dots, M\}$ of stochastic models with respective parameter spaces Θ_m embedded into Euclidean spaces of various dimensions. The joint Bayesian distribution sets

(i) a prior on the model space $\mathscr{M}$, $\pi(1), \dots, \pi(M)$,

(ii) for each model m, a prior on its parameter space Θ_m, whose density with respect to a reference measure (often the Lebesgue measure of the Euclidean space) is $\pi_m(\theta_m)$ and

(iii) the likelihood of the data $\mathbf{Y}$ within each model, namely $\pi_m(\mathbf{y} \mid \theta_m)$.

Consider the extended parameter space $\Theta = \bigcup_{m=1}^{M} \{m\} \times \theta_m$, the Bayesian analysis targets posterior model probabilities, that is the marginal in $\mathscr{M}$ of the posterior distribution for $(m, \theta_1, \dots, \theta_M)$ given $\mathbf{Y} = \mathbf{y}$. By Bayes theorem, the posterior probability of model m is

$$\pi(m \mid \mathbf{y}) = \frac{e(\mathbf{y} \mid m)\pi(m)}{\sum_{m'=1}^{M} e(\mathbf{y} \mid m')\pi(m')},$$

where $e(\mathbf{y} \mid m)$ is the evidence of model m defined as

$$e(\mathbf{y} \mid m) = \int_{\Theta_m} \pi_m(\mathbf{y} \mid \theta_m)\pi_m(\theta_m)\mathrm{d}\theta_m. \tag{9.1}$$

When the goal of the Bayesian analysis is the selection of the model that best fits the observed data $\mathbf{y}^{\text{obs}}$, it is performed through the maximum *a posteriori* (MAP) defined by

$$\widehat{m}_{\text{MAP}}(\mathbf{y}^{\text{obs}}) = \arg\max_m \pi(m \mid \mathbf{y}^{\text{obs}}). \tag{9.2}$$

One faces the usual difficulties of Markov random fields to compute the posterior model distribution $\pi(m \mid \mathbf{y}^{\text{obs}})$. In the hidden case the problem is even more complicated than parameter estimation issues and can be termed as a triply-intractable problem. Indeed the stochastic model for $\mathbf{Y}$ is based on the latent process $\mathbf{X}$ in $\mathscr{X}$, that is

$$\pi_m(\mathbf{y} \mid \theta_m) = \int_{\mathscr{X}} \pi(\mathbf{y} \mid \mathbf{x}, \phi_m)\pi(\mathbf{x} \mid \psi_m, \mathscr{G}_m)\mu(\mathrm{d}\mathbf{x}), \tag{9.3}$$

with μ the counting measure (discrete case) or the Lebesgue measure (continuous case). Both the integral and the Gibbs distribution are intractable and consequently so is the posterior distribution.

9.2 ABC model choice approximation

Approximate Bayesian computation (ABC) is a simulation based approach that offers a way to circumvent the difficulties of models which are intractable but can be simulated from. Subsequently to a work of Tavaré et al. (1997) in population genetics, the method is introduced by Pritchard et al. (1999) as a genuine acceptance-rejection method (see Algorithm 7). The basis is to sample from an approximation of the target distribution (8.1), namely

$$\pi_\epsilon\left(\psi \,\middle|\, \mathbf{y}^{\text{obs}}, \mathscr{G}\right) \propto \int_{\mathscr{Y}} \pi(\psi)\pi\left(\mathbf{y} \,\middle|\, \psi, \mathscr{G}\right) K_\epsilon\left(\mathbf{y} \,\middle|\, \mathbf{y}^{\text{obs}}\right) \mathrm{d}\mathbf{y},$$

where $K_\epsilon(\cdot \mid \mathbf{y}^{\text{obs}})$ is a probability density on the configuration space $\mathscr{Y}$ centered on $\mathbf{y}^{\text{obs}}$ with a support defined by ϵ. In its original version, assuming a metric space $(\mathscr{Y}, \rho)$, this density is set

to the uniform distribution on the ball $\mathscr{B}(\epsilon, \mathbf{y}^{\text{obs}})$ of radius ϵ centered at $\mathbf{y}^{\text{obs}}$, that is

$$K_\epsilon\left(\mathbf{y} \,\middle|\, \mathbf{y}^{\text{obs}}\right) \propto \mathbf{1}\left\{\mathbf{y} \in \mathscr{B}\left(\epsilon, \mathbf{y}^{\text{obs}}\right)\right\} = \mathbf{1}\left\{\rho\left(\mathbf{y}, \mathbf{y}^{\text{obs}}\right) \leq \epsilon\right\}.$$

The use of a kernel function instead of the latter has been studied by Wilkinson (2013). Concerning the calibration of ϵ, a trade-off has to be found to ensure good performances of the procedure. If the threshold is small enough, $\pi_\epsilon(\cdot \mid \mathbf{y}^{\text{obs}}, \mathscr{G})$ provides an accurate approximation that may nonetheless suffer from a high computational cost. For the limiting case $\epsilon = 0$, we recover the true posterior distribution. However decreasing the threshold, while maintaining the amount of simulations accepted, can be problematic in terms of processing time since the acceptance probability can be too low, if not zero, *i.e.*, $\mathbf{P}(\rho(\mathbf{Y}, \mathbf{y}^{\text{obs}}) \leq \epsilon) = \int_{\mathscr{Y}} \pi(\mathbf{y} \mid \psi, \mathscr{G}) \mathbf{1}\left\{\rho(\mathbf{y}, \mathbf{y}^{\text{obs}}) \leq \epsilon\right\} \mathrm{d}\mathbf{y} \to 0$. Conversely, a large threshold ϵ leads to a poor approximation of the posterior distribution since almost all simulated particles are accepted, *i.e.*, $\lim_{\epsilon \to \infty} \mathbf{P}(\rho(\mathbf{Y}, \mathbf{y}^{\text{obs}}) \leq \epsilon) = 1$. The standard solution is to pick out an empirical quantile of the distance (*e.g.*, Beaumont et al., 2002). We refer the reader to Marin et al. (2012) and the reference therein for an overview of this calibration question. This point is also discussed further in Chapter **??**.

Algorithm 7: Acceptance-rejection algorithm

Input: an observation $\mathbf{y}^{\text{obs}}$, summary statistics S, a number of iterations T, an empirical quantile of the distance T_ϵ

Output: a sample from the approximated target of $\pi_\epsilon\left(\cdot \,\middle|\, \mathbf{y}^{\text{obs}}, \mathscr{G}\right)$

for $t \leftarrow 1$ **to** T **do**
  draw ψ from $\pi(\cdot)$;
  draw $\mathbf{y}$ from $\pi\left(\cdot \,\middle|\, \psi, \mathscr{G}\right)$;
  compute $\mathbf{S}(\mathbf{y})$;
  save $\left\{\psi^{(t)}, \mathbf{S}\left(\mathbf{y}^{(t)}\right)\right\} \leftarrow \left\{\psi, \mathbf{S}\left(\mathbf{y}\right)\right\}$;
end
sort the replicates according to the distance $\rho\left\{\mathbf{S}\left(\mathbf{y}^{(t)}\right), \mathbf{S}\left(\mathbf{y}^{\text{obs}}\right)\right\}$;
keep the T_ϵ first replicates;
return the sample of accepted particles

In practical terms, the data usually lies in a space of high dimension and the algorithm faces the curse of dimensionality, namely that is almost impossible to sample dataset in the neighbourhood of $\mathbf{y}$. The ABC algorithm performs therefore a (non linear) projection of the observed and simulated datasets onto some Euclidean space of reasonable dimension d via a function s, composed of summary statistics. The use of summary statistics in place of the data leads to the

pseudo-target

$$\pi_\epsilon\left(\psi \,\middle|\, \mathbf{S}\left(\mathbf{y}^{\text{obs}}\right), \mathscr{G}\right) \propto \int_{\mathscr{Y}} \pi(\psi)\pi\left(\mathbf{y} \,\middle|\, \psi, \mathscr{G}\right) \mathbf{1}\left\{\rho\left(\mathbf{S}(\mathbf{y}), \mathbf{S}\left(\mathbf{y}^{\text{obs}}\right)\right) \leq \epsilon\right\} \mathrm{d}\mathbf{y}.$$

Beyond the seldom situation where s is sufficient, *i.e.*, $\mathbf{P}\left(\psi \mid s(\mathbf{y}^{\text{obs}})\right) = \mathbf{P}(\psi \mid \mathbf{y}^{\text{obs}})$, we cannot recover better than $\pi\left(\psi \mid \rho\left\{s(\mathbf{y}), s(\mathbf{y}^{\text{obs}})\right\} \leq \epsilon\right)$. Hence the calibration of ABC can become complicated due to the difficulty or even the impossibility to quantify the effect of the different approximations. Recent articles have proposed automatic schemes to construct these statistics (rarely from scratch but based on a large set of candidates) for Bayesian parameter inference and are meticulously reviewed by Blum et al. (2013) who compare their performances in concrete examples.

Example (Curse of dimensionality)**.** *Consider* $\mathbf{Y}, \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(N)}$ *a sequence of random variables in* $\mathbb{R}^d$ *independent and identically distributed according to the uniform distribution on* $[0,1]^d$*. Denote* $d_\infty(d,N)$ *the distance function to* $\mathbf{Y}$ *defined as*

$$d_\infty(d,N) = \mathbf{E}\left\{\min_{i=1,\ldots,N} \|\mathbf{Y} - \mathbf{Y}^{(i)}\|_\infty\right\},$$

where $\|\cdot\|_\infty$ *stands for the supremum norm on* $\mathbb{R}^d$*.*

$$\begin{aligned} d_\infty(d,N) &= \int_0^\infty \mathbf{P}\left(\min_{i=1,\ldots,N} \|\mathbf{Y} - \mathbf{Y}^{(i)}\|_\infty > t\right) \mathrm{d}t \\ &= \int_0^\infty 1 - \mathbf{P}\left(\min_{i=1,\ldots,N} \|\mathbf{Y} - \mathbf{Y}_{(i)}\|_\infty \leq t\right) \mathrm{d}t. \end{aligned}$$

Due to the independence assumption, the latter can be written as

$$\begin{aligned} \mathbf{P}\left(\min_{i=1,\ldots,N} \|\mathbf{Y} - \mathbf{Y}_{(i)}\|_\infty \leq t\right) &\leq N\mathbf{P}\left(\|\mathbf{Y} - \mathbf{Y}^{(1)}\|_\infty \leq t\right) \\ &\leq N(2t)^d \end{aligned}$$

Starting from $1 - N(2t)^d \geq 0$ *for* $t \leq \left(2N^{1/d}\right)^{-1}$*, we get the following lower bound*

$$d_\infty(d,N) \geq \int_0^{\left(2N^{1/d}\right)^{-1}} 1 - N(2t)^d \mathrm{d}t = \frac{d}{2(d+1)} N^{-\frac{1}{d}}.$$

Table 1 yields the lower bound for various dimension space d *and sample sizes* N*. The latter shows how paramount the calibration of the threshold* ϵ *is. When dealing with discrete Markov random field, the dimension of* $\mathscr{Y}$ *is* $K^{|\mathscr{S}|} = K^n$*, that is for binary random variables defined on a* 10×10 *lattice the dimension of the configuration space is* $2^{100} \approx 10^{30}$*.*

Table 1: Illustration of the curse of dimensionality for various dimension d and sample sizes N.

$d_\infty(d,N)$	$N=100$	$N=1000$	$N=10000$	$N=100000$
$d_\infty(1,N)$	0.0025	0.00025	0.000025	0.0000025
$d_\infty(2,N)$	≥ 0.033	≥ 0.01	≥ 0.0033	≥ 0.001
$d_\infty(10,N)$	≥ 0.28	≥ 0.22	≥ 0.18	≥ 0.14
$d_\infty(200,N)$	≥ 0.48	≥ 0.48	≥ 0.47	≥ 0.46

Once the parameter space includes models index $\mathscr{M}$, the ABC model choice follows the same vein than the above ABC methodology used for Bayesian parameter inference. To approximate $\widehat{m}_{\text{MAP}}$, ABC starts by simulating numerous triplets $(m,\theta_m,\mathbf{y})$ from the joint Bayesian model. Afterwards, the algorithm mimics the Bayes classifier (9.2): it approximates the posterior probabilities by the frequency of each model number associated with simulated $\mathbf{y}$'s in a neighbourhood of $\mathbf{y}^{\text{obs}}$. If required, we can eventually predict the best model with the most frequent model in the neighbourhood, or, in other words, take the final decision by plugging in (9.2) the approximations of the posterior probabilities.

At this stage, this first, naive algorithm faces the curse of dimensionality illustrated in Example 9.2. Then the algorithm compares the observed data $\mathbf{y}^{\text{obs}}$ with numerous simulations $\mathbf{y}$ through summary statistics $\mathbf{S}(\cdot)=\left\{s_1(\cdot),\ldots,s_M(\cdot)\right\}$, that is the concatenation of the summary statistics of each models with cancellation of possible replicates.

Algorithm 8: ABC model choice algorithm

Input: an observation $\mathbf{y}^{\text{obs}}$, summary statistics $\mathbf{S}$, a number of iterations T, an empirical quantile of the distance T_ϵ

Output: a sample from the approximated target of $\pi_\epsilon\left(\cdot \mid \mathbf{S}(\mathbf{y}^{\text{obs}}),\mathscr{G}\right)$

for $t \leftarrow 1$ **to** T **do**
  draw m from π;
  draw θ from π_m;
  draw $\mathbf{y}$ from $\pi_m(\cdot \mid \theta)$;
  compute $\mathbf{S}(\mathbf{y})$;
  save $\left\{m^{(t)},\psi^{(t)},\mathbf{S}\left(\mathbf{y}^{(t)}\right)\right\} \leftarrow \left\{m,\psi,\mathbf{S}\left(\mathbf{y}\right)\right\}$;
end
sort the replicates according to the distance $\rho\left(\mathbf{S}\left(\mathbf{y}^{(t)}\right),\mathbf{S}\left(\mathbf{y}^{\text{obs}}\right)\right)$;
keep the T_ϵ first replicates;
return the sample of accepted particles

The accepted particles $(m^{(t)},\mathbf{y}^{(t)})$ at the end of Algorithm 8 are distributed according to $\pi(m \mid$

$\rho\left(\mathbf{S}(\mathbf{y}), \mathbf{S}(\mathbf{y}^{\text{obs}})\right) \leq \epsilon$) and the estimate of the posterior model distribution is given by

$$\hat{\pi}_\epsilon\left(m \,\middle|\, \mathbf{y}^{\text{obs}}\right) = \frac{\sum \mathbf{1}\left\{m^{(t)} = m, \rho\left(\mathbf{S}\left(\mathbf{y}^{(t)}\right), \mathbf{S}\left(\mathbf{y}^{\text{obs}}\right)\right) \leq \epsilon\right\}}{\sum \mathbf{1}\left\{\rho\left(\mathbf{S}\left(\mathbf{y}^{(t)}\right), \mathbf{S}\left(\mathbf{y}^{\text{obs}}\right)\right) \leq \epsilon\right\}}.$$

The choice of such summary statistics presents major difficulties that have been especially highlighted for model choice (Robert et al., 2011, Didelot et al., 2011). When the summary statistics are not sufficient for the model choice problem, Didelot et al. (2011) and Robert et al. (2011) found that the above probability can greatly differ from the genuine $\pi(m \mid \mathbf{y}^{\text{obs}})$.

Model selection between Markov random fields whose energy function is of the form $H(\mathbf{y} \mid \theta, \mathscr{G}) = \theta^T s(\mathbf{y})$, such as the Potts model, is a surprising example for which ABC is consistent. Indeed Grelaud et al. (2009) have pointed out that the exponential family structure ensures that the vector of summary statistics $\mathbf{S}(\cdot) = \left\{s_1(\cdot), \ldots, s_M(\cdot)\right\}$ is sufficient for each model but also for the joint parameter across models $(\mathscr{M}, \theta_1, \ldots, \theta_M)$. This allows to sample exactly from the posterior model distribution when $\epsilon = 0$. However the fact that the concatenated statistic inherits the sufficiency property from the sufficient statistics of each model is specific to exponential families (Didelot et al., 2011). When dealing with model choice between hidden Markov random fields, we fall outside of the exponential families due to the bound to the data. Thus we face the major difficulty outlined by Robert et al. (2011): it is almost impossible to build a sufficient statistic of reasonable dimension, *i.e.*, of dimension much lower than the dimension of $\mathscr{X}$.

Beyond the seldom situations where sufficient statistics exist and are explicitly known, Marin et al. (2014) provide conditions which ensure the consistency of ABC model choice. The present dissertation has thus to answer the absence of available sufficient statistics for hidden Potts fields as well as the difficulty (if not the impossibility) to check the above theoretical conditions in practice. If much attention has been devoted to Bayesian parameter inference (*e.g.*, Blum et al., 2013), very few has been accomplished in the context of ABC model choice apart from the work of Prangle et al. (2014). The statistics $\mathbf{S}(y)$ reconstructed by Prangle et al. (2014) have good theoretical properties (those are the posterior probabilities of the models in competition) but are poorly approximated with a pilot ABC run (Robert et al., 2011), which is also time consuming.

9.3 Bayesian Information Criterion approximations

In most cases, we could not design good summary statistics for ABC model choice. The method thus implies a loss of statistical information and raises many questions from the choice of summary statistics to the consistency of the algorithm. This makes the implementation of the pro-

cedure particularly difficult, the use of the whole dataset being impossible due to the curse of dimensionality. In place of a fully Bayesian approach, model choice criterion can be used.

As presented in Section 9.1, the Bayesian approach to model selection is based on posterior model probabilities. Under the assumption of model being equally likely *a priori*, the posterior model distribution writes as

$$\pi(m \mid \mathbf{y}) = \frac{e(\mathbf{y} \mid m)}{\sum_{m'=1}^{M} e(\mathbf{y} \mid m')}.$$

Hence, the MAP rule (9.2) is equivalent to choose the model with the largest evidence (9.1). The integral is usually intractable, thus much of the research in model selection area focuses on evaluating it by numerical methods.

The Bayesian Information Criterion (BIC) is a simple but reliable solution to approximate the evidence using Laplace method (Schwarz, 1978, Kass and Raftery, 1995). It corresponds to the maximized log-likelihood with a penalization term, namely

$$\mathrm{BIC}(m) = -2\log\pi_m\left(\mathbf{y} \,\middle|\, \hat{\theta}_{\mathrm{MLE}}\right) + d_m \log(n) \approx -2\log\pi(\mathbf{y} \mid m), \tag{9.4}$$

where $\hat{\theta}_{\mathrm{MLE}}$ is the maximum likelihood estimate for $\pi_m(\mathbf{y} \mid \theta_m)$, d_m is the number of free parameters of model m (usually the dimension of Θ_m) and $n = |\mathscr{S}|$ is the number of sites. The model with the highest posterior probability is the one that minimizes BIC. The criterion is closely related to the Akaike Information Criterion (AIC, Akaike, 1973) that solely differs in the penalization term:

$$\mathrm{AIC}(m) = -2\log\pi_m\left(\mathbf{y} \,\middle|\, \hat{\theta}_{\mathrm{MLE}}\right) + 2d_m.$$

AIC has been widely compared to BIC (*e.g.*, Burnham and Anderson, 2002). Looking at the penalization term indicates than BIC tends to favor simpler models than those picked by AIC. We shall also mention that AIC has been shown to overestimate the number of parameters, even asymptotically (*e.g.*, Katz, 1981). We refer the reader to Kass and Raftery (1995) and the references therein for a more detailed discussion on AIC.

BIC is an asymptotic estimate of the evidence whose error is bounded as the sample size grows to infinity regardless of the prior π_m on the parameter space (Schwarz, 1978), see Chapter **??** for a more detailed presentation. The approximation may seem somewhat crude due to this

$\mathcal{O}(1)$ error. However as observed by Kass and Raftery (1995) the criterion does not appear to be qualitatively misleading as long as the sample size n is much larger than the number d_m of free parameters in the model.

This dissertation tackles the issue of selecting a number of components from a collection of hidden Markov random fields. The use of BIC might be questionable due to the absence of results on the reliability of the evidence estimate in this context. Though we follow an argument of Forbes and Peyrard (2003) that arises from the work of Gassiat (2002) in hidden Markov chains.

> *"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, Gassiat (2002) has proven 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."*

Difficulties in the context of hidden Markov random field are of two kinds and both come from the maximized log-likelihood term $\log \pi_m(\mathbf{y} \mid \hat{\theta}_{\text{MLE}})$. Neither the maximum likelihood estimate $\hat{\theta}_{\text{MLE}}$ (see Section 6.2) nor the incomplete likelihood (9.3) are available since they would require to integrate a Gibbs distribution over the latent space configuration. As regards the simpler case of observed Markov random field solutions have been brought by penalized pseudolikelihood (Ji and Seymour, 1996) or MCMC approximation of BIC (Seymour and Ji, 1996). Over the past decade, only few works have addressed the model choice issue for hidden Markov random field from that BIC perspective. Arguably the most relevant has been suggested by Forbes and Peyrard (2003) who, among other things, generalize an earlier approach of Stanford and Raftery (2002). Their proposal is to use mean field-like approximations introduced in Section 6.2 to estimate BIC. But other attempts based on simulations techniques have been investigated (Newton and Raftery, 1994). Regarding the question of inferring the number of latent states, one might avocate in favor of the Integrated Completed Likelihood (ICL, Biernacki et al., 2000). This opportunity has been explored by Cucala and Marin (2013) but their complex algorithm cannot be extended easily to choose the dependency structure.

Approximations of the Gibbs distribution

The central question is the evaluation of the incomplete likelihood (9.3), that is

$$\pi_m\left(\mathbf{y} \mid \theta_m\right) = \sum_{\mathbf{x} \in \mathscr{X}} \pi\left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi_m\right) \pi\left(\mathbf{x} \,\middle|\, \psi_m, \mathscr{G}_m\right).$$

The most straightforward approach to circumvent the computational burden is to replace the Gibbs distribution with some simpler distributions such as the mean-field like approximations (see Section 6.2), namely

$$\pi\left(\mathbf{x} \,\middle|\, \psi_m, \mathscr{G}\right) \approx \mathbf{P}^{\text{MF-like}}\left(\mathbf{x} \,\middle|\, \psi_m, \mathscr{G}\right) = \prod_{i \in \mathscr{S}} \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}_{\mathscr{N}(i)}, \psi_m, \mathscr{G}\right). \tag{9.5}$$

The latter corresponds to an incomplete likelihood estimate of the form

$$\mathbf{P}_m^{\text{MF-like}}\left(\mathbf{y} \,\middle|\, \theta_m\right) = \prod_{i \in \mathscr{S}} \sum_{\mathbf{x}_i} \pi\left(y_i \mid x_i, \phi_m\right) \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}_{\mathscr{N}(i)}, \psi_m, \mathscr{G}\right).$$

This results in the following approximation of BIC

$$\text{BIC}^{\text{MF-like}}(m) = -2 \log \mathbf{P}_m^{\text{MF-like}}\left(\mathbf{y} \,\middle|\, \hat{\theta}_{\text{MLE}}\right) + d_m \log(n). \tag{9.6}$$

This approach includes the Pseudolikelihood Information Criterion (PLIC) of Stanford and Raftery (2002) as well as the mean field-like approximations of BIC proposed by Forbes and Peyrard (2003). For the latter, the authors suggest to use for $(\tilde{\mathbf{x}}, \hat{\theta}_{\text{MLE}})$ the output of the VEM-like algorithm based on the mean-field like approximations described in Section 6.2. As regards neighbourhood restoration step, Forbes and Peyrard (2003) advocate in favor of the simulated field algorithm (see Algorithm 5).

Stanford and Raftery (2002) suggest to approximate the Gibbs distribution in (9.3) with the pseudolikelihood of Qian and Titterington (1991). Note the latter differs from the pseudolikelihood of Besag (1975). Instead of integrating over $\mathscr{X}$, the idea is to consider as $\tilde{\mathbf{x}}$ a configuration close to the Iterated Conditional Modes (ICM, Besag, 1986) estimate of $\mathbf{x}$. ICM is an iterative procedure that aims at finding an estimate of

$$\mathbf{x}_{\text{MAP}} = \arg\max_{\mathbf{x}} \pi\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta, \mathscr{G}\right).$$

In its unsupervised version it alternates between a restoration step of the latent states and an estimation step of the parameter θ. The restoration step corresponds to a sequential update of the sites, namely given the current configuration $\tilde{\mathbf{x}}^{(t)}$ and the current parameter $\theta^{(t)}$

$$\tilde{x}_i^{(t+1)} = \arg\max_{x_i} \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathcal{N}(i)} = \tilde{\mathbf{x}}_{\mathcal{N}(i)}^{(t)}, \mathbf{y}, \hat{\theta}^{(t)}, \mathcal{G}\right).$$

Afterwards the parameter is updated given the new configuration $\tilde{\mathbf{x}}^{(t+1)}$, the spatial component being updated by maximizing the pseudolihood (5.3),

$$\begin{aligned}
\phi^{(t+1)} &= \arg\max_{\phi} \log \pi\left(\mathbf{y} \,\middle|\, \tilde{\mathbf{x}}^{(t+1)}, \phi\right), \\
\psi^{(t+1)} &= \arg\max_{\psi} \log f_{\text{pseudo}}\left(\tilde{\mathbf{x}}^{(t+1)} \mid \psi, \mathcal{G}\right).
\end{aligned}$$

Denote $\left(\mathbf{x}^{\text{ICM}}, \theta^{\text{ICM}}\right)$ the output of the ICM algorithm, PLIC can be written as

$$\begin{aligned}
\text{PLIC}(m) = -2\log\Bigg\{ \prod_{i \in \mathcal{S}} \sum_{\mathbf{x}_i} \pi\left(y_i \mid x_i, \phi_m^{\text{ICM}}\right) \\
\pi\left(x_i \,\middle|\, \mathbf{X}_{\mathcal{N}(i)} = \mathbf{x}_{\mathcal{N}(i)}^{\text{ICM}}, \psi_m^{\text{ICM}}, \mathcal{G}\right) \Bigg\} + d_m \log(n). \quad (9.7)
\end{aligned}$$

Stanford and Raftery (2002) have also proposed the Marginal Mixture Information Criterion (MMIC) but for the latter they report less satisfactory results.

Approximation of the partition function

Forbes and Peyrard (2003) have also derived another criterion considering that BIC can express only in terms of partition functions. Let $Z(\psi, \mathcal{G})$ and $Z(\theta, \mathcal{G})$ denote the respective normalizing constants of the latent and the conditional fields (see Section 2.4), namely,

$$\begin{aligned}
Z\left(\psi, \mathcal{G}\right) &= \sum_{\mathbf{x} \in \mathcal{X}} \exp\left\{-H\left(\mathbf{x} \,\middle|\, \psi, \mathcal{G}\right)\right\}, \\
Z\left(\theta, \mathcal{G}\right) &= \sum_{\mathbf{x} \in \mathcal{X}} \exp\left\{-H\left(\mathbf{x} \,\middle|\, \mathbf{y}, \phi, \psi, \mathcal{G}\right)\right\} = \sum_{\mathbf{x} \in \mathcal{X}} \pi\left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi\right) \exp\left\{-H\left(\mathbf{x} \,\middle|\, \psi, \mathcal{G}\right)\right\}.
\end{aligned}$$

Starting from the Bayes formula, the incomplete likelihood can be written as

$$\pi\left(\mathbf{y} \,\middle|\, \theta\right) = \frac{\pi\left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi\right) \pi\left(\mathbf{x} \,\middle|\, \psi, \mathcal{G}\right)}{\pi\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta, \mathcal{G}\right)} = \frac{\pi\left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi\right) \exp\left\{-H\left(\mathbf{x} \,\middle|\, \psi, \mathcal{G}\right)\right\}}{\exp\left\{-H\left(\mathbf{x} \,\middle|\, \mathbf{y}, \phi, \psi, \mathcal{G}\right)\right\}} \frac{Z\left(\theta, \mathcal{G}\right)}{Z\left(\psi, \mathcal{G}\right)}$$

which using the definition of the Hamiltonian $H(\mathbf{x} \mid \mathbf{y}, \phi, \psi, \mathcal{G})$ simplifies into

$$\pi\left(\mathbf{y} \mid \theta\right) = \frac{Z\left(\theta, \mathcal{G}\right)}{Z\left(\psi, \mathcal{G}\right)}.$$

The expression (9.4) turns into

$$\mathrm{BIC}(m) = -2\log Z\left(\theta, \mathcal{G}\right) + 2\log Z\left(\psi, \mathcal{G}\right) + d_m \log(n).$$

Hence, the problem of estimating the Gibbs distribution becomes a problem of estimating the normalizing constants. The latter issue could be addressed with Monte Carlo estimator such as the path sampling (Gelman and Meng, 1998) while being time consuming. Forbes and Peyrard (2003) propose to use instead a first order approximation of the normalizing constant arising from mean field theory.

Consider $\mathbf{P}^{\mathrm{MF}}(\cdot \mid \psi, \mathcal{G})$ the mean field approximation of the Gibbs distribution $\pi(\cdot \mid \psi, \mathcal{G})$. The mean field approximation can be written as follows

$$\mathbf{P}^{\mathrm{MF}}\left(\mathbf{x} \mid \psi, \mathcal{G}\right) = \frac{1}{Z^{\mathrm{MF}}\left(\psi, \mathcal{G}\right)} \exp\left\{-H^{\mathrm{MF}}\left(\mathbf{x} \mid \psi, \mathcal{G}\right)\right\},$$

where $Z^{\mathrm{MF}}(\psi, \mathcal{G})$ and $H^{\mathrm{MF}}(\mathbf{x} \mid \psi, \mathcal{G})$ are the mean field expressions for the normalizing constant and the Hamiltonian. It is worth repeating that the mean field approximation is the minimizer of the Kullback-Leibler divergence over the set of probability distributions that factorize and hence both quantities are easy to compute. Denote $\mathbf{E}^{\mathrm{MF}}$ the expectation under the mean field approximation, the Kullback-Leibler divergence can be written as

$$\mathrm{KL}\left(\mathbf{P}^{\mathrm{MF}}(\cdot \mid \psi, \mathcal{G}), \pi(\cdot \mid \psi, \mathcal{G})\right) = \mathbf{E}^{\mathrm{MF}}\left(\log\left\{\frac{\mathbf{P}^{\mathrm{MF}}(\mathbf{X} \mid \psi, \mathcal{G})}{\pi(\mathbf{X} \mid \psi, \mathcal{G})}\right\}\right).$$

It follows from the positivity of the Kullback-Leibler divergence

$$Z\left(\psi, \mathcal{G}\right) \geq Z^{\mathrm{MF}}\left(\psi, \mathcal{G}\right) \exp\left(-\mathbf{E}^{\mathrm{MF}}\left\{H\left(\mathbf{X} \mid \psi, \mathcal{G}\right) - H^{\mathrm{MF}}\left(\mathbf{X} \mid \psi, \mathcal{G}\right)\right\}\right). \tag{9.8}$$

The mean field approximation yields the optimal lower bound which is used as an estimate of the normalizing constant. The same applies to the Gibbs distribution $\pi(\cdot \mid \mathbf{y}, \theta, \mathcal{G})$ and we denote

$Z^{\mathrm{MF}}(\theta,\mathscr{G})$ and $H^{\mathrm{MF}}(\cdot \mid \mathbf{y},\theta,\mathscr{G})$ the corresponding mean field expressions for the normalizing constant and the Hamiltonian. It follows another approximation of BIC, namely

$$\begin{aligned}\mathrm{BIC}^{\mathrm{GBF}}(m) = &-2\log\left\{Z^{\mathrm{MF}}\left(\hat{\theta}_m^{\mathrm{MLE}},\mathscr{G}\right)\right\} + 2\log\left\{Z^{\mathrm{MF}}\left(\hat{\psi}_m^{\mathrm{MLE}},\mathscr{G}\right)\right\} \\ &+ 2\mathbf{E}^{\mathrm{MF}}\left\{H\left(\mathbf{X}\,\middle|\,\mathbf{y},\hat{\theta}_m^{\mathrm{MLE}},\mathscr{G}\right) - H^{\mathrm{MF}}\left(\mathbf{X}\,\middle|\,\mathbf{y},\hat{\theta}_m^{\mathrm{MLE}},\mathscr{G}\right)\right\} \\ &- 2\mathbf{E}^{\mathrm{MF}}\left\{H\left(\mathbf{X}\,\middle|\,\hat{\psi}_m^{\mathrm{MLE}},\mathscr{G}\right) - H^{\mathrm{MF}}\left(\mathbf{X}\,\middle|\,\hat{\psi}_m^{\mathrm{MLE}},\mathscr{G}\right)\right\} \\ &+ d_m \log(n). \end{aligned} \tag{9.9}$$

Forbes and Peyrard (2003) argue that the latter is more satisfactory than $\mathrm{BIC}^{\text{MF-like}}(m)$ in the sense it is based on a optimal lower bound for the normalizing constants contrary to the mean field-like approximations. However that does not ensure better results as regards model selection.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akademinai Kiado, 1973.

M. Alfò, L. Nieddu, and D. Vicari. A finite mixture model for image segmentation. *Statistics and Computing*, 18(2):137–150, 2008.

C. Andrieu and G. O. Roberts. The Pseudo-Marginal Approach for Efficient Monte Carlo Computations. *The Annals of Statistics*, 37(2):697–725, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

A. Barbu and S.-C. Zhu. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1239–1253, 2005.

M. A. Beaumont. Estimation of Population Growth or Decline in Genetically Monitored Populations. *Genetics*, 164(3):1139–1160, 2003.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035, 2002.

J. E. Besag. Nearest-neighbour Systems and the Auto-Logistic Model for Binary Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(1):75–83, 1972.

J. E. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

J. E. Besag. Statistical Analysis of Non-Lattice Data. *The Statistician*, 24:179–195, 1975.

J. E. Besag. On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.

J. E. Besag and P. J. Green. Spatial Statistics and Bayesian Computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):25–37, 1993.

J. E. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000.

M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. *Statistical Science*, 28(2):189–208, 2013.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002.

A. Caimo and N. Friel. Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55, 2011.

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

G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003.

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

P. Clifford. Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, pages 19–32, 1990.

C. Cooper and A. M. Frieze. Mixing properties of the Swendsen-Wang process on classes of graphs. *Random Structures and Algorithms*, 15(3-4):242–261, 1999.

L. Cucala and J.-M. Marin. Bayesian Inference on a Mixture Model With Spatial Dependence. *Journal of Computational and Graphical Statistics*, 22(3):584–597, 2013.

L. Cucala, J.-M. Marin, C. P. Robert, and D. M. Titterington. A Bayesian Reassessment of Nearest-Neighbor Classification. *Journal of the American Statistical Association*, 104(485):263–273, 2009.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

X. Descombes, R. D. Morris, J. Zerubia, and M. Berthod. Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood. *Image Processing, IEEE Transactions on*, 8(7):954–963, 1999.

X. Didelot, R. G. Everitt, A. M. Johansen, and D. J. Lawson. Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):49–76, 2011.

R. G. Edwards and A. D. Sokal. Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical review D*, 38(6):2009, 1988.

R. G. Everitt. Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks. *Journal of Computational and Graphical Statistics*, 21(4):940–960, 2012.

F. Forbes and G. Fort. Combining Monte Carlo and Mean Field-Like Methods for Inference in Hidden Markov Random Fields. *Image Processing, IEEE Transactions on*, 16(3):824–837, 2007.

F. Forbes and N. Peyrard. Hidden Markov random field model selection criteria based on mean field-like approximations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1089–1101, 2003.

O. François, S. Ancelet, and G. Guillot. Bayesian Clustering Using Hidden Markov Random Fields in Spatial Population Genetics. *Genetics*, 174(2):805–816, 2006.

O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395): 832–842, 1986.

N. Friel. Bayesian Inference for Gibbs Random Fields Using Composite Likelihoods. In *Proceedings of the Winter Simulation Conference*, number 28 in WSC '12, pages 1–8. Winter Simulation Conference, 2012.

N. Friel and A. N. Pettitt. Likelihood Estimation and Inference for the Autologistic Model. *Journal of Computational and Graphical Statistics*, 13(1):232–246, 2004.

N. Friel and H. Rue. Recursive computing and simulation-free inference for general factorizable models. *Biometrika*, 94(3):661–672, 2007.

N. Friel, A. N. Pettitt, R. Reeves, and E. Wit. Bayesian Inference in Hidden Markov Random Fields for Binary Data Defined on Large Lattices. *Journal of Computational and Graphical Statistics*, 18(2):243–261, 2009.

E. Gassiat. Likelihood ratio inequalities with applications to various mixtures. In *Annales de l'IHP Probabilités et statistiques*, volume 38, pages 897–906, 2002.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, 13(2):163–185, 1998.

S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

S. Geman and C. Graffigne. Markov Random Field Image Models and Their Applications to Computer Vision. In *Proceedings of the International Congress of Mathematicians*, volume 1, pages 1496–1517, 1986.

H. Georgii. *Gibbs Measures and Phase Transitions*. De Gruyter studies in mathematics. De Gruyter, 2011.

C. J. Geyer. Markov Chain Monte Carlo Maximum Likelihood. 1991.

C. J. Geyer and E. A. Thompson. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):657–699, 1992.

V. K. Gore and M. R. Jerrum. The Swendsen–Wang process does not always mix rapidly. *Journal of Statistical Physics*, 97(1-2):67–86, 1999.

P. J. Green and S. Richardson. Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association*, 97(460):1055–1070, 2002.

A. Grelaud, C. P. Robert, J.-M. Marin, F. Rodolphe, and J.-F. Taly. ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 4(2):317–336, 2009.

G. R. Grimmett. A theorem about random fields. *Bulletin of the London Mathematical Society*, 5(1):81–84, 1973.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

D. M. Higdon. Discussion on the Meeting on the Gibbs Sampler and Other Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B*, 55(1):78, 1993.

D. M. Higdon. Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association*, 93(442):585–595, 1998.

M. Huber. A bounding chain for Swendsen-Wang. *Random Structures & Algorithms*, 22(1):43–59, 2003.

M. A. Hurn, O. K. Husby, and H. Rue. A Tutorial on Image Analysis. In *Spatial Statistics and Computational Methods*, volume 173 of *Lecture Notes in Statistics*, pages 87–141. Springer New York, 2003.

E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik*, 31:253–258, 1925.

C. Ji and L. Seymour. A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *The annals of applied probability*, pages 423–443, 1996.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430): 773–795, 1995.

R. W. Katz. On Some Criteria for Estimating the Order of a Markov Chain. *Technometrics*, 23(3): 243–249, 1981.

I. Lanford, O.E. and D. Ruelle. Observables at infinity and states with short range correlations in statistical mechanics. *Communications in Mathematical Physics*, 13(3):194–215, 1969.

J. S. Liu. Peskun's theorem and a modified discrete-state gibbs sampler. *Biometrika*, 83(3):681–682, 1996.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian Computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

J.-M. Marin, N. S. Pillai, C. P. Robert, and J. Rousseau. Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):833–859, 2014.

V. Matveev and R. Shrock. Complex-temperature singularities in Potts models on the square lattice. *Physical Review E*, 54(6):6174, 1996.

C. A. McGrory, D. M. Titterington, R. Reeves, and A. N. Pettitt. Variational Bayes for estimating the parameters of a hidden Potts model. *Statistics and Computing*, 19(3):329–340, 2009.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of Chemical Physics*, 21(6):1087–1092, 1953.

A. Mira, J. Møller, and G. O. Roberts. Perfect slice samplers. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(3):593–606, 2001.

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

M. T. Moores, C. E. Hargrave, F. Harden, and K. Mengersen. Segmentation of cone-beam CT using a hidden Markov random field with informative priors. *Journal of Physics : Conference Series*, 489, 2014.

I. Murray, Z. Ghahramani, and D. J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press, 2006.

R. Neal and G. Hinton. A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In M. Jordan, editor, *Learning in Graphical Models*, volume 89 of *NATO ASI Series*, pages 355–368. Springer Netherlands, 1998.

M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48, 1994.

L. Onsager. Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition. *Phys. Rev.*, 65:117–149, 1944.

A. N. Pettitt, N. Friel, and R. W. Reeves. Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(1):235–246, 2003.

R. B. Potts. Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, volume 48, pages 106–109. Cambridge Univ Press, 1952.

D. Prangle, P. Fearnhead, M. P. Cox, P. J. Biggs, and N. P. French. Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, 13(1):67–82, 2014.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

J. G. Propp and D. B. Wilson. Exact Sampling with Coupled Markov chains and Applications to Statistical Mechanics. *Random structures and Algorithms*, 9(1-2):223–252, 1996.

W. Qian and D. Titterington. Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337 (1647):407–428, 1991.

R. Reeves and A. N. Pettitt. Efficient recursions for general factorisable models. *Biometrika*, 91 (3):751–757, 2004.

C. P. Robert, J.-M. Cornuet, J.-M. Marin, and N. S. Pillai. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108 (37):15112–15117, 2011.

G. O. Roberts and J. S. Rosenthal. Convergence of slice sampler Markov chains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):643–660, 1999.

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

H. Rue. Fast Sampling of Gaussian Markov Random Fields. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2):325–338, 2001.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.

T. Rydén and D. Titterington. Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*, 7(2):194–211, 1998.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

L. Seymour and C. Ji. Approximate Bayes model selection procedures for Gibbs-Markov random fields. *Journal of Statistical Planning and Inference*, 51(1):75–97, 1996.

D. C. Stanford and A. E. Raftery. Approximate Bayes factors for image segmentation: The pseudolikelihood information criterion (PLIC). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(11):1517–1520, 2002.

R. H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2):86–88, 1987.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring Coalescence Times From DNA Sequence Data. *Genetics*, 145(2):505–518, 1997.

S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. *Psychometrika*, 61(3):401–425, 1996.

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

R. D. Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2): 129–141, 2013.

U. Wolff. Collective Monte Carlo updating for spin systems. *Physical Review Letters*, 62(4):361, 1989.

C. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1): 95–103, 1983.

F.-Y. Wu. The Potts model. *Reviews of modern physics*, 54(1):235, 1982.

L. Younes. Estimation and annealing for Gibbsian fields. *Annales de l'Institut Henri Poincaré*, 24:269–294, 1988.

J. Zhang. The mean field theory in EM procedures for Markov random fields. *Signal Processing, IEEE Transactions on*, 40(10):2570–2583, 1992.