Méthodes d'inférence statistique pour champs de Gibbs

Julien Stoehr

Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université de Montpellier**

Préparée au sein de l'école doctorale **I2S - Information, Structures, Systèmes**
Et de l'unité de recherche **UMR 5149 - IMAG - Institut Montpelliérain Alexander Grothendieck**

Spécialité: **Biostatistique**

Présentée par **Julien Stoehr**

Méthodes d'inférence statistique pour champs de Gibbs

Soutenue le 29 octobre 2015 devant le jury composé de

Jean-Michel Marin	Professeur	Université de Montpellier	Directeur
Pierre Pudlo	Professeur	Université de Montpellier	Co-directeur
Lionel Cucala	Maître de Conférences	Université de Montpellier	Co-encadrant
Florence Forbes	Directeur de recherche	INRIA Grenoble Rhône-Alpes	Rapporteur
Håvard Rue	Professeur	Norwegian University of Science and Technology	Rapporteur
Stéphanie Allassonnière	Professeur chargée de cours	École Polytechnique	Examinateur
Stéphane Robin	Directeur de recherche	INRA/AgroParisTech	Président du jury

Université de Montpellier

Do the best you can until you know better.
Then, when you know better, do better.
— Maya Angelou

À la mémoire de ma grand-mère Marie-Madeleine, de mon grand-père Paul
et de mon amie Christiane.

Remerciements

S'il est un exercice qui s'avère tout aussi difficile que la rédaction de la thèse, c'est bien l'écriture des remerciements. J'espère qu'au travers de ces lignes, je n'oublierai de rendre hommage à aucune des personnes qui m'ont permis d'en être arrivé là.

Mes premiers remerciements vont tout naturellement à mes directeurs de thèse Jean-Michel Marin, Pierre Pudlo et Lionel Cucala. Jean-Michel pour m'avoir initié à la statistique bayésienne, pour avoir su laisser s'épanouir ma liberté de recherche tout en me proposant un encadrement sans faille. Pour ton enthousiasme, pour ton soutien constant aussi bien professionnel que personnel. Et enfin pour m'avoir fait découvrir les beautés du vignoble languedocien. Pierre pour ta patience et ta disponibilité face à l'étudiant chronophage et plein de questions que j'ai pu être. Pour ton investissement hors norme qui m'a permis de gagner en maturité et de prendre du recul sur mes travaux. Pour tes enseignements qui ont grandement contribué aux compétences que j'ai développées au cours de cette thèse, soient elles mathématiques ou informatiques. Lionel pour avoir été l'acteur de certaines tâches ingrates comme traiter à la main des centaines de données. Pour ton œil de lynx infaillible qui traquait sans relâche mes coquilles, parfois, plus que nombreuses. À tous les trois, pour la confiance que vous m'avez accordée et pour tout ce que je ne peux simplement résumer en quelques lignes, merci.

Je voudrais ensuite exprimer toute ma gratitude à Florence Forbes et Håvard Rue[1] qui m'ont fait l'honneur de rapporter cette thèse avec beaucoup d'attention et de rigueur malgré des emplois du temps chargés. L'intérêt que vous avez porté à mes travaux, nos discussions et vos remarques sont pour moi le meilleur encouragement à poursuivre dans le monde de la recherche.

À Stéphanie Allassonnière et Stéphane Robin pour avoir accepté de faire partie de ce jury sans la moindre hésitation. J'aimerais, ici, également remercier Nathalie Peyrard.

[1] Jeg vil gjerne takke Håvard Rue, som gjorde meg den ære å gjennomgå denne avhandlingen på en oppmerksom og grundig måte til tross for en travel timeplan. Interessen du har vist for mitt arbeid, og i diskusjoner og tilbakemelding, er for meg den beste motivasjonen for å fortsette å forske.

Ce fut pour moi un immense privilège que Stéphanie et toi ayez suivi cette thèse depuis ses débuts. Vos conseils lors des comités de suivi de thèse ont joué un rôle primordial dans la réussite de ces travaux.

De façon plus générale, je souhaite remercier les membres de l'IMAG, et tout particulièrement ceux de l'équipe EPS, pour votre accueil chaleureux. Un énorme merci à Sophie Cazanave-Pin, Bernadette Lacan, Myriam Debus, Éric Hugounenq et Nathalie Quintin. De par votre gentillesse, votre disponibilité, votre sympathie et votre aide précieuse en toute circonstance, vous êtes indiscutablement les super-héros du bâtiment 9. Une petite pensée pour Gemma Bellal qui a toujours un mot gentil. À Christian, Catherine, Élodie, André, Fabien, Irène, Benjamin, Christophe, Gwladys, Jean-Noël, Benoîte, Xavier, Ludovic, Cyrille, Baptiste et Vanessa, que ce soit de près ou de loin, vous avez contribué à faire de cette thèse une expérience enrichissante et plaisante.

Une mention toute spéciale à Mathieu Ribatet pour tous nos échanges mathématiques, politiques ou sportifs autour d'une bonne bière (ou d'un plateau du restaurant administratif... au choix). Ton amitié et ton soutien signifient beaucoup. Tu as accompagné mes premiers pas d'enseignant et m'as fait profiter de tes compétences aussi bien pédagogiques que scientifiques. Et si je devais te faire une dédicace particulière dans cette thèse, ce serait le chapitre 2 auquel tu as grandement contribué sans être officiellement impliqué.

Ali, je ne t'oublie pas. Tu me pardonneras les approximations orthographiques mais que ce soit pour nos discussions sérieuses ou grivoiseries : choukran jazilan !

À celui qui en plus d'être mathématicien est également un SwingJammerz... Simon, merci pour ton humour et ta simplicité.

I would like to thank Nial Friel. Your enthusiasm as I was just beginning my thesis has been a source of motivation. Our collaboration is an integral part of this dissertation and has driven us into interesting questions. I am looking forward working with you in the future and developing all the project we already have.

Parmi tous mes professeurs, que je salue, je souhaite rendre un hommage tout particulier à deux d'entre eux. Nathalie Baumgarten, c'est avec toi que j'ai appris à lire. Tu es le point de départ de ces longues années d'études et savoir qu'après tout ce temps, tu continues de suivre mes pas d'écolier me fait chaud au cœur. Puis, cette aventure montpelliéraine n'aurait pas eu lieu sans Benoît Cadre. Tu es celui par qui je suis arrivé à la statistique. Tes conseils avisés et tes enseignements ont rendu tout cela possible et je t'en suis extrêmement reconnaissant.

Mes mots s'adressent maintenant aux doctorants actuels et passés pour les discus-

sions délirantes à base de petits poneys ou autres extravagances, et pour ces moments de procrastination savamment dosés. Petit clin d'œil à mes co-bureaux : Benjamin qui m'aura appris que les vraies mathématiques ne sont que des flèches à la craie blanche sur un tableau noir... Et Joubine, que dire si ce n'est que tu resteras mon interlocuteur privilégié concernant nordpresse.be ! Tu sais le succès que je te souhaite. Faire une thèse, c'est aussi adopter une deuxième famille. Un énorme merci à mon grand-frère de thèse Mohammed, ma petite sœur de thèse Coralie et mon petit demi-frère (ça commence à devenir compliqué la filiation) Paul pour les moments de craquage complet et tout le reste. Un seul regret : Momo il va définitivement falloir que tu danses avec nous à la SFDS ! Merci à Mickaël Lallouche, partner in crime avec qui j'ai pris énormément de plaisir à organiser le séminaire des doctorants. La liste est encore longue et je n'ai pas de place ici pour raconter une anecdote sur tous mais vous avez chacun eu une place particulière pendant cette thèse. Donc merci à Mathieu, Angelina, Christophe, Yousri (aka Yousri the King), Étienne, Myriam, Jérémie, Gautier, Nahla, Théo, Quentin, Boushra, Nejib, Romain, Wallid, Amina, Emmanuelle, Claudia, Alexandre, Hanen, Guillaume, David, Mike, Antoine, Alaaeddine, Anis, Jocelyn, Rebecca, Anthony, Wenran, Samuel, Francesco, Christian, Tutu, Elsa.

À cette longue liste je souhaite ajouter toutes les personnes avec qui j'ai pu discuter lors de congrès ou de séminaires. Chaque échange a été l'occasion de faire avancer mes travaux, de parfaire ou d'enrichir ma culture scientifique. Un merci tout particulier à Christian Robert pour ses critiques toujours bénéfiques, à Benjamin Guedj pour toutes ses recommandations, et enfin à Florian Maire pour son aide à Dublin.

L'achèvement d'une thèse passe par le subtil dosage entre ambiance studieuse et sas de décompression. Tout ce travail n'aurait donc simplement pas été possible sans ma famille et mes amis. À commencer par la dream team qui a répondu présente à chacune de mes convocations loufoques. Je ne compte plus les matchs joués ni les buts marqués, mais Arnold (tu peux le dire maintenant que tu as 28 ans...), Romain, Rémi, Fabien, Isaie, Will, Ju, Tom, Max et Nico : une chose est sûre, ce qu'on a toujours réussi ce sont les troisièmes mi-temps.

Je me dois ensuite de remercier ceux qui ont partagé une bonne partie de ma vie pendant cette thèse : les SwingJammerz. À Oliv, Ben dit Scapino, JB, Tat pour m'avoir accueilli et inspiré au quotidien. À Isaie et Claire pour votre confiance et votre soutien permanent. À Sand, Mimi, Chacha, et Christelle, pour avoir été des partenaires et des amies exceptionnelles. À Richard, Mélody, Billy, Annabelle et Sixte pour m'avoir fait découvrir les joies de la team dont je ne peux citer le nom ici ! À Anne pour m'avoir transmis le virus irlandais. À Karine pour ton optimisme indéfectible. À tous, un énorme merci pour les centaines de danses et les bons moments que nous avons

partagés. À mes chouchous qui m'ont suivi et supporté en cours de danse comme en dehors, les Joséphine (Myriam, Émilie, Sally), David, Franck, Clara, Jean-Marie, Marion & Marion, Jérémy et Fabien : keep on swinging !

À tous mes amis de la scène qui ont impulsé mes travaux sur le parquet comme en dehors : Paulo, JP, Marlène, Irène, Lionel, Audrey, Aurélien, Rija, Peter, Marielle, Béné, Paulopaul, Liliane, Henry, Margaux, Alexis, Aline, Niko, Audrey, Patrick, Patricia, Jean-Marc, Céline, Séb, Anne-Laure et tout ceux que j'oublie de citer. À Charlie mon camarade de voyage. À Flo une amie, une confidente, une révélatrice de créativité. À Line, Gildas, Maxence et Corentin, l'alliance parfaite de l'Alsace et de la Bretagne. À Thomas, Annie, Max, Dax et Sarah pour m'avoir aidé à me dépasser, à persévérer ! À Will et Maéva pour avoir été là dans les bons comme dans les mauvais moments, pour me rappeler au quotidien que tout ou presque est possible à force de travail. Je vous embrasse !

À cette liste quasi exhaustive, il faut ajouter celles et ceux qui ont su faire de Montpellier un foyer chaleureux où il fait bon vivre. À mes collocs Mathieu, Maïwenn et Jérémie qui ont fait en sorte que la rédaction de ce manuscrit se passe dans des conditions idéales. Votre amitié a été plus que bénéfique et appréciable dans les derniers kilomètres de ce marathon qu'est la thèse. J'ai également une pensée pour mes collocs des premiers jours Alice et Peter. De L.A. à Montpellier vous avez tous deux apporté beaucoup de soleil dans ma vie de doctorant. À Katy, une muse inattendue qui a su faire avancer ma réflexion. À Sally et Sylvia pour avoir donné une toute autre dimension aux soirées Love Boat. À Hélian qui me laisse pantois d'admiration. À Julien mon bricoleur préféré, tu as été un compagnon de galère et le complice parfait en toute circonstance. À Vincent, mein elsässicher Freund, le seul ici à connaître la véritable grandeur de notre région... Et surtout le seul à savoir que le Sundgau n'est pas une région d'Afrique. À Pacchus mon président de cœur plein de philosophie. À Jean et Maud parce qu'on pourrait presque tout résumer à une soirée quizz épique. À Leslie, Ev' Ma, Val et Pauline, pour les bouffées d'air frais. À Pierre Malaka pour l'analyse anthropologique et psychologique des microcosmes ambiants ou tout simplement pour être toi. À Carine un amour de volleyeuse pour ton oreille attentive et ta joie de vivre.

Pour terminer, il me reste à rendre hommage à ceux qui sont les plus chers à mon cœur. À ma marraine et René, mon parrain et Léa, Fabienne, Didier, Lulu, Nico, Alex, Élise, Patrice, Marie-Christine, Christian et Natacha pour avoir été présents, pour m'avoir encouragé d'aussi loin que je m'en souvienne. À la Bes Léa, au Bof Guillaume et à vos parents pour être une belle famille au top niveau. À Bernard, tu as été pour moi un peu comme un grand-père... Avec Christiane vous n'avez manqué aucune des étapes importantes qui m'ont conduit ici. Cela a toujours compté énormément pour moi et

mon seul regret aujourd'hui est qu'elle ne soit pas avec nous pour partager ce moment. À Mamie, même dépassée par mon parcours universitaire tu t'es toujours souciée de mon avenir. Le petit garçon qui courait dans les prés et les forêts d'Horodberg a bien grandi et j'aurais aimé que Papi voit ce que l'homme que je suis devenu a réussi à accomplir. À Mémé qui malgré la distance m'a toujours porté, m'a toujours donné la motivation pour faire de mon mieux. Je sais que tu continueras à veiller sur moi de là où tu es. À Noémie, Sébastien et Baptiste, j'espère remplir mon rôle de grand-frère aussi bien que possible. Il n'existe pas qu'un seul exemple de réussite et vous en êtes l'exemple parfait. J'espère vous apporter au quotidien autant que ce que chacun d'entre vous m'apporte. À Maman et Papa vous m'avez enseigné le respect, l'humilité et l'amour du travail bien fait. Même dans les moments de doutes, vous n'avez jamais cessé de croire en moi. J'ai toujours eu à cœur d'exceller pour être à la hauteur des sacrifices que, sans hésiter, vous avez faits pour nous, pour être digne de l'amour et du soutien que vous nous avez toujours témoignés. La plus belle réussite, le plus beau cadeau, c'est de vous savoir fiers et, en dépit de tous les aléas de la vie, inconditionnellement présents. À tous, simplement et sincèrement merci !

Contents

List of Figures

List of Tables

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 neighbors, 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. 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), these models have enjoyed great success in image analysis (*e.g.*, Stanford and Raftery, 2002, Celeux et al., 2003, Forbes and Peyrard, 2003, Hurn et al., 2003, Alfò et al., 2008, Moores et al., 2014) but also in other applications including disease mapping (*e.g.*, Green and Richardson, 2002) and genetic analysis (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).

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 Bayesian inference as the computation of the likelihood is an integral part of the procedure. 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. Mention first of all likelihood approximations involving a product of easily normalised distributions: the pseudolikelihood (Besag, 1974, 1975), mean field approximations (*e.g.*, Celeux et al., 2003, Forbes and Peyrard, 2003), the reduced dependence approximation (Friel et al., 2009) and composite likelihoods (*e.g.*, Okabayashi et al., 2011, Friel, 2012). On the other hand Monte Carlo approaches have played a major role to estimate the intractable likelihood such as the maximum likelihood estimator of Geyer and Thompson (1992) or the path sampling approach of Gelman and Meng (1998). More recently Møller et al. (2006) present an auxiliary variable scheme that tackles this problem by cancelling out the estimation of the normalizing constant, a work then further developed by Murray et al. (2006) in their exchange algorithm. Another opportunity is the approximate Bayesian computation (Pritchard et al., 1999) which provides a Monte Carlo approximation of the targeted distribution. These manifold techniques are reviewed among others and compared by Everitt (2012). Their main drawback is the computing time involved that can be considerable. Alternatively McGrory et al. (2009) construct a variational Bayes scheme with more efficient computing time to analyse hidden Potts model.

The present work cares about the problem of carrying out Bayesian inference for Markov random field. When dealing with hidden random fields, the focus is solely on hidden data represented by Ising or 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 $\mathbf{x}$.

Adjustment of posterior parameter distribution approximations

The first part of the present dissertation proposes to adjust substitute to the intractable posterior parameter distribution. One of the earliest approaches to overcome the troublesome constant is the pseudolikelihood method (Besag, 1974, 1975) which replaces the likelihood function by the product of tractable full-conditional distributions of all nodes. However the pseudolikelihood is not a genuine probability distribution and leads to unreliable estimate of parameters (*e.g.*, Geyer, 1991, Friel and Pettitt, 2004, Cucala et al., 2009). Despite this drawback, the pseudolikelihood has been used,

if only for its simplicity of calculation, in a wide range of applications, especially in hidden Markov settings following the work of Besag et al. (1991) like in Heikkinen and Hogmander (1994), Rydén and Titterington (1998). A natural generalization to consider is the composite likelihood (Lindsay, 1988) which refines pseudolikelihood by considering products of larger collections of variables. Composite likelihood has been made popular in a context where marginal distributions can be computed (*e.g.*, Varin et al., 2011, and references therein). Spatial lattice processes differ from that class of models in the sense that dependence structure makes impossible the calculation of marginal probabilities and require the application of conditional composite likelihood instead.

The purpose of this work is to use such composite likelihood methods for Bayesian inference on observed Markov random fields. Currently, there is very little literature on that possibility, although Pauli et al. (2011) and Ribatet et al. (2012) present a discussion on the use of marginal composite likelihoods in a Bayesian setting. As the neighbors relationship for Potts model is too complicated, the interest is solely on conditional composite likelihood. Friel (2012) had a similar focus and studied how the size of the collections of variables influences the resulting approximate posterior distribution. His work follows a study conducted by Okabayashi et al. (2011) although from a likelihood inference perspective. As in this dissertation, both consider composite likelihood consisting in a product of joint distributions of collections of neighbouring variables, namely blocks (or windows) of the lattice. The peculiarity of Friel (2012) lies in the exact computation of conditional composite likelihood for moderately large blocks using the recursive algorithm of Reeves and Pettitt (2004) for general factorizable models, like Potts model, a method generalizing a result known for hidden Markov models (Zucchini and Guttorp, 1991, Scott, 2002). The latter recursion is a tempting alternative to stochastic approximations such as the path sampling approach (Gelman and Meng, 1998). In the same way of Friel et al. (2009), we plug it in a procedure that leads to reliable estimates based on exact calculation on small lattices.

The main contribution of this work is the adjustment of posterior distributions resulting from using a misspecified likelihood function, referred to as composite posterior distribution. Indeed, Friel (2012) is interested in the impact of the size of the blocks, but he does not take advantage of the possibly weighting of blocks, even though he observes that non-calibrated composite likelihood leads to overly precise posterior parameters due to a substantial lower variability of the surrogate distribution. The adjustment of composite likelihoods has long-standing antecedents in the frequentist paradigm (*e.g.*, Geys et al., 2001, Chandler and Bate, 2007), the primary goal being to recover a chi-squared asymptotic null distribution. Our approach differs from the

latter in the sense we make a shift from asymptotical behaviour of the composite likelihood to local matching conditions about the posterior distribution. A key feature to our proposal is the closed form of the gradient and of the Hessian matrix of the log-posterior. Consequently it is possible to implement optimization algorithms that allows to adjust the mode and the curvature at the mode of the composite posterior distribution. Note that similar approach has been proposed in the context of Gaussian Markov random fields (Rue et al., 2009).

Here we focus especially on how to formulate conditional composite likelihoods for application to the autologistic model with possible anisotropy on the lattice. We present numerical result for lattices small enough so that the true posterior distribution can be computed using the recursion of Reeves and Pettitt (2004) and serves as a ground truth against which to compare the adjusted and non-adjusted composite posterior distributions. The calibration is achieved for composite likelihoods that use exhaustively all the blocks of the image, a challenging situation since there is a multiple use of the data. The good results make this procedure an option worth exploring for more complex settings such as hidden data or exponential random graph.

Approximate Bayesian computation model choice between hidden Markov random fields

The second part of the current work aims at addressing the problem of selecting a dependency structure for a hidden Markov random field in the Bayesian paradigm and explores the opportunity of approximate Bayesian computation (*e.g.,* Tavaré et al., 1997, Pritchard et al., 1999, Marin et al., 2012, Baragatti and Pudlo, 2014). Up to our knowledge, this important question has not yet been addressed in the Bayesian literature. Alternatively we could have tried to set up a reversible jump Markov chain Monte Carlo, but follows an important work for the statistician to adapt the general scheme, as shown by Caimo and Friel (2011, 2013) in the context of exponential random graph models where the observed data is a graph.

The Bayesian approach to model selection is based on posterior model probabilities. When dealing with models whose likelihood cannot be computed analytically, Bayesian model choice becomes challenging since the evidence of each model writes as the integral of the likelihood over the prior distribution of the model parameter. To answer the question of model choice, different opportunities have been tackled in the literature but approximate Bayesian computation (ABC) method has appeared as one of the most satisfactory approach to deal with intractable likelihood . ABC is a simulation based approach that compares the observed data $\mathbf{y}^{\text{obs}}$ with numerous

simulations $\mathbf{y}$ through summary statistics $\mathbf{S}(\mathbf{y})$ in order to supply a Monte Carlo approximation of the posterior probabilities of each model. 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). Beyond the seldom situations where sufficient statistics exist and are explicitly known (Gibbs random fields are surprising examples, see Grelaud et al., 2009), Marin et al. (2014) provide conditions which ensure the consistency of ABC model choice. The present work 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.

Recent articles have proposed automatic schemes to construct theses 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. But very few has been accomplished in the context of ABC model choice apart from the work of Prangle et al. (2014). The statistics $\mathbf{S}(\mathbf{y})$ 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.

ABC model choice is here presented as a k-nearest neighbor classifier, and we define a local error rate which is the first contribution of the current work. We also provide an adaptive ABC algorithm based on the local error to select automatically the dimension of the summary statistics. The second contribution is the introduction of a general and intuitive approach to produce geometric summary statistics for hidden Potts model. This part of the dissertation concludes with numerical results in that framework. Especially, we show with our approach that the number of simulation required by ABC can be significantly cut down reducing at the same time the computational cost whilst preserving performances.

Approximate model choice criterion: the Block Likelihood Information Criterion

The last contribution considers model choice criterion for selecting the probabilistic model that best accounts for the observation. This work is motivated by a more general issue than the choice of an underlying graph for which we explore the opportunity of the Bayesian Information Criterion (BIC) (Schwarz, 1978) to overcome the computational burden of ABC algorithms. Model choice is a problem of probabilistic model comparison. The standard approach to compare one model against another is based on the Bayes factor (Kass and Raftery, 1995) that involves the ratio of the

evidence of each model. As already mentioned the evidence can not be computed with standard procedure due to a high-dimensional integral. Various approximations have been proposed but a commonly used one, if only for its simplicity, is BIC that is an asymptotic estimate of the evidence based on the Laplace method. The criterion is is a simple penalized function of the maximized log-likelihood. In this last part of the dissertation, we provide an approximation of BIC able to infer both the number of latent states and the dependency structure of a discrete hidden Markov random field. The question of inferring the number of latent states has been recently tackled by Cucala and Marin (2013) with an Integrated Completed Likelihood criterion(Biernacki et al., 2000) but their complex algorithm cannot be extended easily to choose the dependency structure.

In the context of Markov random fields, the difficulty comes from the maximized log-likelihood part in BIC. Indeed, it involves the Gibbs distribution whose exact computation is generally not feasible. For observed random field solutions proposed to circumvent the problem are based for example on penalized pseudolikelihood (Ji and Seymour, 1996) and MCMC approximations of BIC (Seymour and Ji, 1996). When the random field is hidden little has been done before the work of Stanford and Raftery (2002) and Forbes and Peyrard (2003). Both are interested in the question of inferring the number of latent states and propose approximations that consist in replacing the true likelihood with a product distribution on system of independent variables to make the computation tractable. Stanford and Raftery (2002) handle the burdensome likelihood with the pseudolikelihood of Qian and Titterington (1991) to yield the so called Pseudo-Likelihood Information Criterion (PLIC). The latter appears to be encompassed in the class of mean field-like approximations of BIC proposed by Forbes and Peyrard (2003).

The proposal of Forbes and Peyrard (2003) derives from variational method that provides a way to approximate the distribution through the introduction of a simpler function that minimizes the Kullback-Leibler divergence between surrogate functions and the Gibbs distribution. The divergence is minimized over the set of probability distributions that factorize in a product on a set of single independent variables. Our main contribution is to show that larger collections of variables, namely blocks of the lattice, can be considered by taking advantage of the exact recursion of Reeves and Pettitt (2004) and leads to an efficient criterion : the Block Likelihood Information Criterion (BLIC). In particular, we will show that a reasonable approximation of the Gibbs distribution is a product of Gibbs distributions on each independent block. To assess the performances of the novel criterion, it is compared to the previous ones on simulated data sets. Overall the criterion shows good results with notable benefits for the estimation of the number of latent states. We fill in our study with a comparison

between BLIC and our second contribution related to ABC.

Overview

Chapter 1 is a reminder on Markov random fields introducing the notation and the model of interest. It is an opportunity to present a brief state of the art related to the inference issues tackled in this dissertation. Each chapter is then dedicated to my own contributions to the analysis of Markov random fields. Chapter 2 presents a correction of composite likelihoods to approximate the posterior distribution of model parameter when the Markov random field is observed. My proposal is based on the modification of the mode and the curvature at the mode of an approximated posterior distribution resulting from a misspecified function to recover the true posterior parameter distribution. This solution is appealing since its computational cost is much lower than the Monte Carlo approaches such as the exchange algorithm or the approximate Bayesian computation. The performances of the correction are illustrated through simulated realizations of isotropic and anisotropic Ising models. Both Chapters 3 and 4 are devoted to the question of model choice between hidden Gibbs random fields. Throughout this dissertation, we tackle two model choice issues: choosing the dependency structure and/or the number of latent states. My first contribution developed in Chapter 3 concerns the approximate Bayesian computation methodology. The major difficulties addressed in Chapter 3 is the absence of relevant summary statistics to choose a latent neighborhood structure. Chapter 3 first introduces a local error rate. The latter aims at evaluating the quality of a set of summary statistics in the absence of the sufficiency property. Then I introduce intuitive geometric summary statistics that leads to an efficient ABC model choice procedure. Some numerical results are given to show the accuracy of the algorithm. Chapter 4 extends the scope to a more general problem: the inference of the number of latent states and the dependency structure. The main contribution of that part is to replace the intractable likelihood with a product distribution on independent blocks of a regular grid of nodes. Contrary to the substitute pixel by pixel proposed in the literature, I suggest to include more spatial information by taking advantage of the opportunity to make exact computation on small enough lattices. This leads a novel approximation of BIC which is compared to PLIC (Stanford and Raftery, 2002) and BIC approximations proposed by Forbes and Peyrard (2003) through simulated data. Conclusions and further discussions are given at the end of the dissertation.

1 Statistical analysis issues for Markov random fields

Markov random fields have been used in many practical settings, surged by the development in the statistical community since the 1970's. 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. The chapter presents a synopsis on the existence of Markov random fields with some specific examples in Section 1.1. The difficulties inherent to the analysis of the stochastic model are especially pointed out. As befits a first chapter, a brief state of the art concerning parameter inference (Section 1.4 and Section 1.5) and model selection (Section 1.8) is presented.

1.1 Markov random field and Gibbs distribution

1.1.1 Gibbs-Markov equivalence

A random field $\mathbf{X}$ is a collection of random variables X_i indexed by a finite set $\mathscr{S} = \{1,\ldots,n\}$, whose elements are called sites, and taking values in a finite state space $\mathscr{X}_i$. In other words $\mathbf{X}$ is a random process on § taking its values in the configuration space $\mathscr{X} = \prod_{i=1}^{n} \mathscr{X}_i$. 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 $\mathscr{S}$.

Markov random fields characterized by local interactions are of special interest. One first introduces an undirected graph $\mathscr{G}$ which induces a topology on $\mathscr{S}$. By definition, sites i and j are adjacent or neighbor if and only if i and j are linked by an edge in $\mathscr{G}$. Denotes $i \overset{\mathscr{G}}{\sim} j$ the adjacency relationship between sites i and j. The neighborhood of site i, denoted hereafter by $\mathscr{N}(i)$, is the set of all the adjacent sites to i in $\mathscr{G}$.

Definition 1. *A random field* $\mathbf{X}$ *is a Markov random field with respect to* $\mathscr{G}$*, if for all*

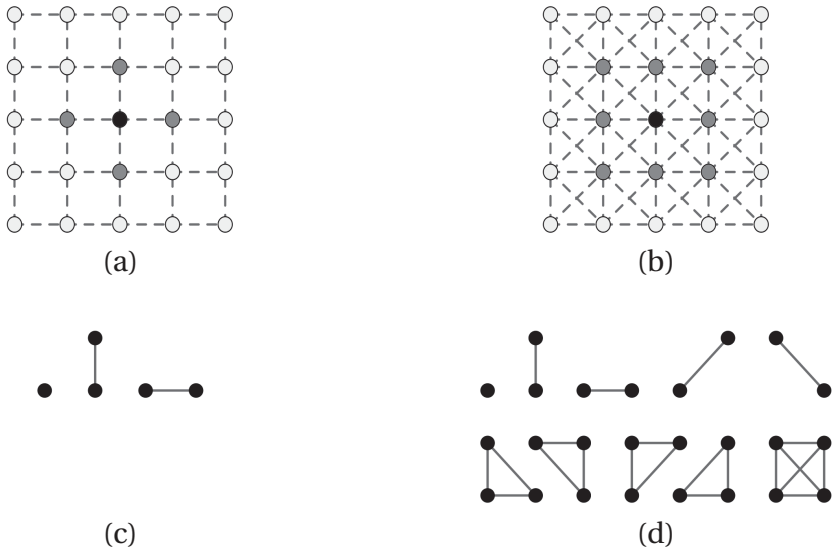


Figure 1.1: First and second order neighborhood graphs $\mathcal{G}$ with corresponding cliques. (a) The four closest neighbors graph $\mathcal{G}_4$. Neighbors of the vertex in black are represented by vertices in gray. (b) The eight closest neighbors graph $\mathcal{G}_8$. Neighbors of the vertex in black are represented by vertices in gray. (c) Cliques of graph $\mathcal{G}_4$. (d) Cliques of graph $\mathcal{G}_8$.

configuration $\mathbf{x}$ *and for all sites* i

$$\mathbf{P}\left(X_i = x_i \mid \mathbf{X}_{-i} = \mathbf{x}_{-i}\right) = \mathbf{P}\left(X_i = x_i \,\middle|\, \mathbf{X}_{\mathcal{N}(i)} = \mathbf{x}_{\mathcal{N}(i)}\right). \tag{1.1}$$

The property (1.1) is a Markov property – the random variable at a site i is conditionally independent of all other sites in $\mathcal{S}$, given its neighbors 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 $\mathcal{G}$ are either the empty set or the entire set of sites $\mathcal{S}$. Recall a clique c in an undirected graph $\mathcal{G}$ is any single vertex or a subset of vertices such that every two vertices in c are connected by an edge in $\mathcal{G}$. However, only Markov random fields with small neighborhood are interesting in practice. Thereafter, we focus on two widely used adjacency structures, namely the graph $\mathcal{G}_4$, respectively $\mathcal{G}_8$, for which the neighborhood 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 1.1. We may speak of first order lattice for $\mathcal{G}_4$ and second order lattice for $\mathcal{G}_8$. The present work makes the analogy with images, such that random variables X_i are shades of grey or colors and the graph $\mathcal{G}$ is a regular grid of pixels.

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. Deriving a consistent joint distribution of a Markov random field through its conditional probabilities is not at all obvious, see Besag (1974) and the references therein. The joint probability is uniquely determined by its conditional probabilities, when it satisfies the positivity condition

$$\mathbf{P}(\mathbf{x}) > 0, \text{ for all configuration } \mathbf{x}. \tag{1.2}$$

Under this assumption, the Hammersley-Clifford theorem yields a characterization of a Markov random field joint probability, namely the distribution of a Markov random field with respect to a graph $\mathscr{G}$ that satisfies the positivity condition (1.2) is a Gibbs distribution for the same topology, see for example Grimmett (1973), Besag (1974) and for a historical perspective Clifford (1990).

Definition 2. *A Gibbs distribution with respect to a graph $\mathscr{G}$ is a probability measure π on $\mathscr{X}$ with the following representation*

$$\pi\left(\mathbf{x} \mid \psi, \mathscr{G}\right) = \frac{1}{Z\left(\psi, \mathscr{G}\right)} \exp\left\{-H\left(\mathbf{x} \mid \psi, \mathscr{G}\right)\right\}, \tag{1.3}$$

where ψ is a free parameter, H denotes the energy function (or Hamiltonian) that decomposes into potential functions V_c associated to the cliques c of $\mathscr{G}$

$$H\left(\mathbf{x} \mid \psi, \mathscr{G}\right) = \sum_c V_c\left(\mathbf{x}_c, \psi\right), \tag{1.4}$$

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

$$Z\left(\psi, \mathscr{G}\right) = \int_{\mathscr{X}} \exp\left\{-H\left(\mathbf{x} \mid \psi, \mathscr{G}\right)\right\} \mu(\mathrm{d}\mathbf{x}). \tag{1.5}$$

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

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

over the set of probability distribution $\mathbf{P}$ on configuration space $\mathscr{X}$ with the same expected energy $\mathbf{E}\left\{H\left(\mathbf{X} \mid \psi, \mathscr{G}\right)\right\} = \int_{\mathscr{X}} H\left(\cdot \mid \psi, \mathscr{G}\right) \mathrm{d}\mathbf{P}$. 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 (1.3), but the joint and the marginal distribution are meanwhile unavailable due to the intractable partition function (1.5). 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.

1.1.2 Autologistic model and related distributions

The Hammersley-Clifford theorem provides valid probability distributions associated with the random variables $X_1, \ldots, X_n$. The formulation in terms of potential allows the local dependency of the Markov field to be specified and leads to a class of flexible parametric models for spatial data. In most cases, cliques of size one (singleton) and two (doubleton) are assumed to be satisfactory to model the spatial dependency and potential functions related to larger cliques are set to zero. Thus, the energy (1.4) becomes

$$H\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \sum_{i=1}^{n} V_i\left(x_i, \alpha\right) + \sum_{i \overset{\mathscr{G}}{\sim} j} V_{ij}\left(x_i, x_j, \beta\right),$$

where $\psi = (\alpha, \beta)$ is the parameter of the Gibbs random field, more precisely α stands for the parameter on sites and β stands for the parameter on edges. The above sum $\sum_{i \overset{\mathscr{G}}{\sim} j}$ ranges the set of edges of the graph $\mathscr{G}$. 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). In what follows, attention is aimed at specific discrete schemes, that is, the space configuration is $\mathscr{X} = \{0, \ldots, K-1\}^n$.

Definition 3. *A Gibbs random field is said to be*

(i) *homogeneous, if the potential V_c is independent of the relative position of the clique c in $\mathscr{S}$,*

(ii) *isotropic, if the potential V_c is independent of the orientation of the clique c.*

Table 1.1: Interaction parametrisation for a homogeneous Gibbs random field in isotropic and anisotropic cases. The table gives values of the parameter β_{ij} corresponding to the orientation of the edge (i, j).

Orientation of edge (i,j)						
Dependency graph	$\mathscr{G}_4$	$\mathscr{G}_8$	$\mathscr{G}_4$	$\mathscr{G}_8$	$\mathscr{G}_8$	$\mathscr{G}_8$
Isotropic Gibbs	β					
Anisotropic Gibbs	β_0		β_1		β_2	β_3

The present dissertation will only focus on homogeneous Markov field with eventual anisotropy. In the anisotropic case, β_{ij} stands for the component of β corresponding to the direction defined by the edge (i, j) but does not depend on the actual position of sites i and j, that is, given two edges (i_1, j_1) and (i_2, j_2) defining the same direction, $\beta_{i_1,j_1} = \beta_{i_2,j_2}$ (see Table 1.1). Mention nevertheless models hereafter do not necessarily impose homogeneity and, indeed, are not tied to a regular lattice.

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

$$\pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \frac{1}{Z\left(\psi, \mathscr{G}\right)} \exp\left\{ \alpha \sum_{i=1}^{n} x_i + \sum_{i \overset{\mathscr{G}}{\sim} j} \beta_{ij} x_i x_j \right\}. \tag{1.6}$$

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 neighbors and themselves observations. The parameter α controls the level of $0-1$ whereas the parameters $\{\beta_{ij}\}$ model the dependency between two neighboring 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 (1.6) 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 neighbors i and j. When $\beta_{ij} > 0$ the interaction is called ferromagnetic and adjacent spins tend to be aligned, that is neighboring 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

$$\pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \frac{1}{Z\left(\psi, \mathscr{G}\right)} \exp\left\{\sum_{i=1}^{n} \sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \sum_{i \overset{\mathscr{G}}{\sim} j} \beta_{ij} \mathbf{1}\{x_i = x_j\}\right\}, \tag{1.7}$$

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 neighboring sites i and j, $\tilde{x}_i \tilde{x}_j = 2\mathbf{1}\{\tilde{x}_i = \tilde{x}_j\} - 1$.

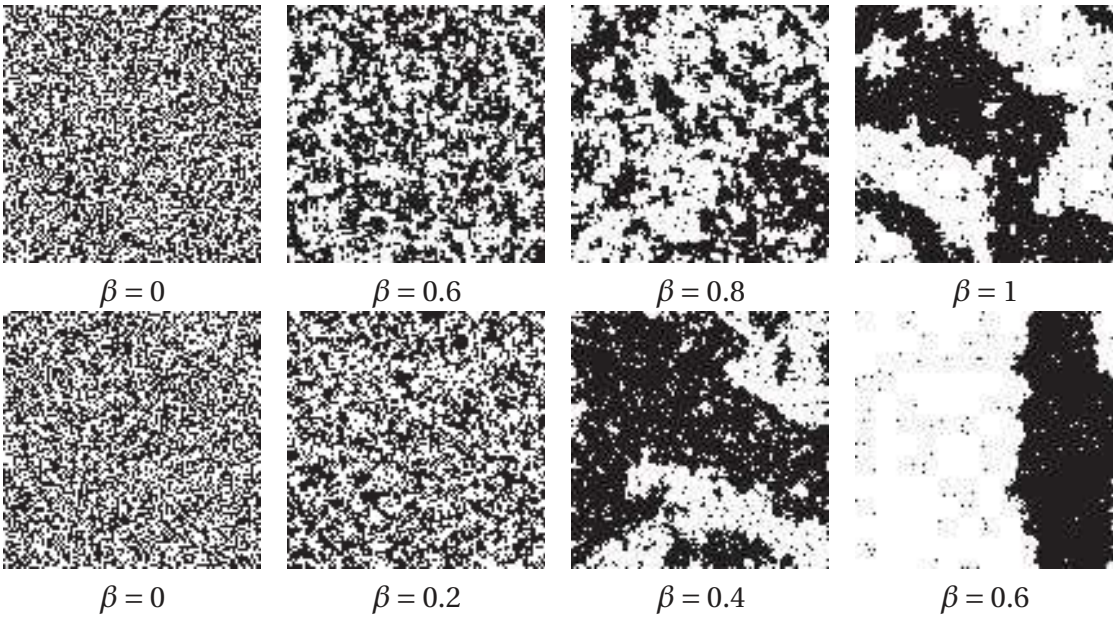


Figure 1.2: Realization of a 2-states Potts model for various interaction parameter β on a 100×100 lattice with a first-order neighborhood (first row) or a second-order neighborhood (second row).

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

$$\underline{\text{Ising}}\text{:}\quad \pi\left(\mathbf{x} \,\middle|\, \beta, \mathscr{G}\right) = \frac{1}{Z\left(\beta, \mathscr{G}\right)} \exp\left\{ \beta \sum_{i \overset{\mathscr{G}}{\sim} j} x_i x_j \right\}, \tag{1.8}$$

$$\underline{\text{Potts}}\text{:}\quad \pi\left(\mathbf{x} \,\middle|\, \beta, \mathscr{G}\right) = \frac{1}{Z\left(\beta, \mathscr{G}\right)} \exp\left\{ \beta \sum_{i \overset{\mathscr{G}}{\sim} j} \mathbf{1}\{x_i = x_j\} \right\}. \tag{1.9}$$

1.1.3 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 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 neighboring sites (see Figure 1.2). 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 neighbors 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.

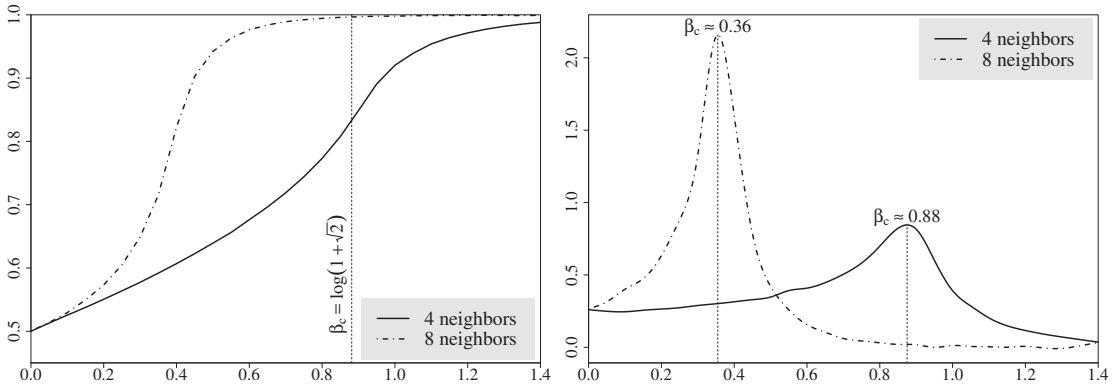
The transition is more rapid than the number of neighbors increases. To illustrate this point, Figure 1.3 gives the average proportion of homogeneous pairs of neighbors, and the corresponding variance, for 2-states Potts model on the first and second order lattices of size 100×100. Indeed, phase transition corresponds to

$$\beta \to \lim_{n\to\infty} \frac{1}{n}\nabla \log Z\left(\beta, \mathscr{G}\right) \text{ is discontinuous at } \beta_c. \tag{1.10}$$

One can show that

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

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



(a) (b)

Figure 1.3: 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 neighbors. (b) Variance of the number of homogeneous pairs of neighbors.

field $\mathbf{X}$, see Section 1.5.1. Condition (1.10) 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.

1.1.4 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., 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 (1.3), 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.*, $\mathbf{y} = (y_i; i \in \mathscr{S})$, 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

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

where $\pi(y_i \mid x_i, \phi)$ 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 neighbors have also been explored (Besag, 1986).

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

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

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

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

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

$$H(\mathbf{x} \mid \mathbf{y}, \phi, \psi, \mathscr{G}) = H(\mathbf{x} \mid \psi, \mathscr{G}) - \sum_{i \in \mathscr{S}} \log \pi(y_i \mid x_i, \phi). \tag{1.11}$$

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

1.2 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.

1.2.1 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).

1.2.2 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 (1.7).

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 neighboring 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

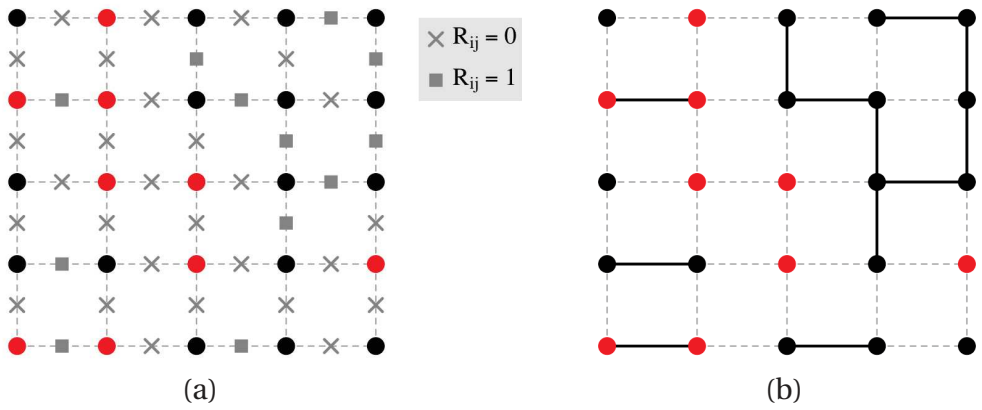


Figure 1.4: Auxiliary variables and subgraph illustrations for the Swendsen-Wang algorithm. (a) Example of auxiliary variables U_{ij} for a 2-states Potts model configuration on the first order square lattice. (b) Subgraph $\Gamma(\mathcal{G}_4, \mathbf{x})$ of the first order lattice $\mathcal{G}_4$ induced by the auxiliary variables U_{ij}.

$\Gamma(\mathcal{G}, \mathbf{x})$ of the graph $\mathcal{G}$ induced by U_{ij} on $\mathbf{x}$, namely the undirected graph made of edges of $\mathcal{G}$ for which $U_{ij} = 1$, see Figure 1.4, two sites belong to the same cluster if and only if there is a path between them in $\Gamma(\mathcal{G}, \mathbf{x})$. Then each cluster $\mathcal{C}$ is assigned to a new state k with probability

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

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.

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 $\mathcal{G}_4$ and $\mathcal{G}_8$ 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

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 1.4`
  Swapping step: assign a new state k to each connected component $\mathcal{C}$ of $\Gamma\left(\mathcal{G}, \mathbf{x}^{(t)}\right)$ with probability $\mathbf{P}\left(\mathbf{X}_{\mathcal{C}}^{(t)} = k\right) \propto \exp\left\{\sum_{i \in \mathcal{C}} \alpha_k\right\}$;
end
return the configuration $\mathbf{x}^{(T)}$

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) that provides exact simulations of Potts models.

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

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

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

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

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 neighborhood system $\mathscr{G}_4$ (see Figure 1.1.(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 neighborhood 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

$$q_i\left(\mathbf{x}_{i:i+h} \,\middle|\, \psi, \mathscr{G}_4\right) = \exp\left(\sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \beta_0 \mathbf{1}\{x_i = x_{i+1}\} + \beta_1 \mathbf{1}\{x_i = x_{i+h}\}\right). \quad (1.12)$$

- When lattice point i is on the last row x_{i+1} drops out ot (1.12), that is

$$q_i\left(\mathbf{x}_{i:i+h} \,\middle|\, \psi, \mathscr{G}_4\right) = \exp\left(\sum_{k=0}^{K-1} \alpha_k \mathbf{1}\{x_i = k\} + \beta_1 \mathbf{1}\{x_i = x_{i+h}\}\right). \quad (1.13)$$

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

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

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 (1.12) and (1.13) 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 neighborhood, factor at a site merely involves its neighbor below and on its right, thereby reducing the number of possible factor to $K^3 + K^2$.

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}

Algorithm 3 presents the scheme I use in my C++ code which is at the core of numerical experiments presented in Chapter 2 and Chapter 4. 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 (1.12) and (1.13) on $\{0, \ldots, K-1\}^3$ and $\{0, \ldots, K-1\}^2$ respectively. Using the previous one-to-one correspondence, the following functions 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

$\{0,\dots,K-1\}\times\{0,\dots,K^h-1\}$

$$q\left(\upsilon_2(k,j)\right) = q\left(k,x_{i+h}\right) = \exp\left(\alpha_k + \beta_1 \mathbf{1}\{x_{i+h}=k\}\right),$$
$$q\left(\upsilon_3(k,j)\right) = q\left(k,x_{i+1},x_{i+h}\right) = \exp\left(\alpha_k + \beta_0 \mathbf{1}\{x_{i+1}=k\} + \beta_1 \mathbf{1}\{x_{i+h}=k\}\right).$$

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

$$q(j) = \exp\left(\sum_{i=n-h+1}^{n}\sum_{k=0}^{K-1}\alpha_k \mathbf{1}\{x_i=k\} + \beta_0 \sum_{i=n-h+1}^{n}\mathbf{1}\{x_i=x_{i+1}\}\right). \tag{1.14}$$

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

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

in Algorithm 3.

It is straightforward to extend this algorithm to hidden Markov random field since as already mention in Section 1.1.4 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 (1.14), 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

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

1.4 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 2.

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,

$$\hat{\phi}_{\mathrm{MLE}} = \arg\max_{\phi} \ \log\pi\left(\mathbf{y} \,\middle|\, \mathbf{x},\phi\right), \tag{1.15}$$

$$\hat{\psi}_{\mathrm{MLE}} = \arg\max_{\psi} \ \log\pi\left(\mathbf{x} \,\middle|\, \psi,\mathscr{G}\right), \tag{1.16}$$

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 (1.15) is straightforward. On the other hand the optimization problem (1.16) 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 (1.5) is the pseudolikelihood (Besag, 1975) which approximates the joint distribution of $\mathbf{x}$ as the product of full-conditional distributions for each site i,

$$f_{\mathrm{pseudo}}\left(\mathbf{x} \,\middle|\, \psi,\mathscr{G}\right) = \prod_{i=1}^{n} \pi\left(x_i \,\middle|\, \mathbf{x}_{-i},\psi,\mathscr{G}\right) = \prod_{i=1}^{n} \frac{\exp\left\{-\sum_{c|i\in c} V_c\left(\mathbf{x}_c,\psi\right)\right\}}{\sum_{\tilde{x}_i} \exp\left\{-\sum_{c|i\in c} V_c\left(\tilde{\mathbf{x}}_c,\psi\right)\right\}}, \tag{1.17}$$

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 neighbors, 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 (1.20), one can show that a unique maximum exists which can be estimated with a simple optimization algorithm.

The pseudolikelihood (1.17) 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 1.5.1 is justified to supersede the maximum pseudolikehood estimate.

1.5 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 2 and Chapter 4.

1.5.1 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 $\psi = (\psi_1, \ldots, \psi_d)$,that is

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

where $\mathbf{S}(\mathbf{x}) = (s_1(\mathbf{x}), \ldots, s_d(\mathbf{x}))$ 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

$$\frac{\partial}{\partial \psi_j} \log Z\left(\psi, \mathscr{G}\right) = \frac{\sum_{\mathbf{x} \in \mathscr{X}} s_j(\mathbf{x}) \exp\left\{\psi^T \mathbf{S}(\mathbf{x})\right\}}{\sum_{\mathbf{x} \in \mathscr{X}} \exp\left\{\psi^T \mathbf{S}(\mathbf{x})\right\}} = \mathbf{E}_\psi\left\{s_j(\mathbf{X})\right\}, \tag{1.18}$$

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

$$\nabla \log \pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \mathbf{S}(\mathbf{x}) - \mathbf{E}_\psi\left\{s(\mathbf{X})\right\}. \tag{1.19}$$

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

$$\nabla^2 \log \pi(\mathbf{x} \mid \psi, \mathscr{G}) = -\mathbf{Var}_\psi\left\{\mathbf{S}(\mathbf{X})\right\}, \tag{1.20}$$

where $\mathbf{Var}_\psi\left\{\mathbf{S}(\mathbf{X})\right\}$ 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 (1.16) is to resort to stochastic approximations on the basis of equation (1.19) (*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.

1.5.2 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 4.

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

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \pi\left(\mathbf{y} \,\middle|\, \theta\right). \tag{1.21}$$

The EM algorithm addresses problem (1.21) 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

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \mathbf{E}\left\{\log \pi\left(\mathbf{X}, \mathbf{y} \,\middle|\, \theta, \mathscr{G}\right) \,\middle|\, \mathbf{Y} = y, \theta^{(t)}\right\} \\
&= \arg\max_{\theta} \sum_{\mathbf{x} \in \mathscr{X}} \pi\left(\mathbf{x} \,\middle|\, \mathbf{y}, \theta, \mathscr{G}\right) \log \pi\left(\mathbf{x}, \mathbf{y} \,\middle|\, \theta, \mathscr{G}\right) \\
&:= \arg\max_{\theta} Q\left(\theta \,\middle|\, \theta^{(t)}\right).
\end{aligned} \tag{1.22}$$

Proposition 1.6. *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 (1.22). 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

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

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

$$\begin{aligned} Q_1\left(\phi \mid \theta^{(t)}\right) &= \mathbf{E}\left\{ \sum_{i\in\mathscr{S}} \log \pi\left(y_i \mid X_i,\phi\right) \,\middle|\, \mathbf{Y}=\mathbf{y},\theta^{(t)} \right\} \\ &= \sum_{i\in\mathscr{S}} \sum_{x_i} \pi\left(x_i \mid \mathbf{y},\theta^{(t)},\mathscr{G}\right) \log \pi\left(y_i \mid x_i,\phi\right), \end{aligned} \tag{1.23}$$

$$\begin{aligned} Q_2\left(\psi \mid \theta^{(t)}\right) &= \mathbf{E}\left\{ -\log Z\left(\psi,\mathscr{G}\right) - \sum_c V_c(\mathbf{X}_c,\psi) \,\middle|\, \mathbf{Y}=\mathbf{y},\theta^{(t)} \right\} \\ &= -\log Z\left(\psi,\mathscr{G}\right) - \sum_c \sum_{\mathbf{x}_c} \pi\left(\mathbf{x}_c \mid \mathbf{y},\theta^{(t)},\mathscr{G}\right) V_c\left(\mathbf{x}_c,\psi\right). \end{aligned} \tag{1.24}$$

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 4 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

$$\log \pi\left(\mathbf{y} \,\middle|\, \theta\right) = \underbrace{\sum_{\mathbf{x}\in\mathscr{X}} \log\left\{\frac{\pi\left(\mathbf{x},\mathbf{y} \,\middle|\, \theta,\mathscr{G}\right)}{\mathbf{P}(\mathbf{x})}\right\}\mathbf{P}(\mathbf{x})}_{=F(\mathbf{P},\theta)} + \underbrace{\sum_{\mathbf{x}\in\mathscr{X}} \log\left\{\frac{\mathbf{P}(\mathbf{x})}{\pi\left(\mathbf{x} \,\middle|\, \mathbf{y},\theta,\mathscr{G}\right)}\right\}\mathbf{P}(\mathbf{x})}_{=\mathrm{KL}\left(\mathbf{P},\, \pi(\cdot|\mathbf{y},\theta,\mathscr{G})\right)}. \quad (1.25)$$

The last KL term denotes the Kullback-Leibler divergence between a given probability distribution $\mathbf{P}$ and the Gibbs distribution $\pi(\cdot \mid \mathbf{y},\theta,\mathscr{G})$. The Kullback-Leibler divergence is a measure of the information lost when one approximates $\pi(\cdot \mid \mathbf{y},\theta,\mathscr{G})$ 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 (1.25) 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 $\pi(\cdot \mid \mathbf{y},\theta,\mathscr{G})$. 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

$$\mathbf{P}^{(t+1)} = \arg\max_{\mathbf{P}\in\mathscr{D}} F\left(\mathbf{P},\theta^{(t)}\right) = \arg\min_{\mathbf{P}\in\mathscr{D}} \mathrm{KL}\left(\mathbf{P},\, \pi\left(\cdot \,\middle|\, \mathbf{y},\theta^{(t)},\mathscr{G}\right)\right), \quad (1.26)$$

$$\theta^{(t+1)} = \arg\max_{\theta} F\left(\mathbf{P}^{(t+1)},\theta\right) = \arg\max_{\theta} \sum_{\mathbf{x}\in\mathscr{X}} \mathbf{P}^{(t+1)}(\mathbf{x}) \log \pi\left(\mathbf{x},\mathbf{y} \,\middle|\, \theta,\mathscr{G}\right). \quad (1.27)$$

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 (1.26) 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 (1.23) and (1.24) are replaced by

$$Q_1^{\text{VEM}}\left(\phi \,\middle|\, \mathbf{P}^{(t)}\right) = \sum_{i \in \mathscr{S}} \sum_{x_i} \mathbf{P}^{(t)}(\mathbf{x}) \log \pi\left(y_i \,\middle|\, x_i, \phi\right), \tag{1.28}$$

$$Q_2^{\text{VEM}}\left(\psi \,\middle|\, \mathbf{P}^{(t)}\right) = -\log Z\left(\psi, \mathscr{G}\right) - \sum_{c} \sum_{\mathbf{x}_c} \mathbf{P}^{(t)}(\mathbf{x}) V_c(\mathbf{x}_c, \psi). \tag{1.29}$$

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 neighbors 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 (1.17) is that neighbors 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

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

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 (1.28) and (1.29) are replaced with

$$\begin{aligned}Q_1^{\text{MF-like}}\left(\phi\,\middle|\,\theta^{(t)}\right)&=\sum_{i\in\mathscr{S}}\sum_{x_i}\pi\left(x_i\,\middle|\,\mathbf{X}_{\mathscr{N}(i)}=\tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)},y_i,\theta^{(t)},\mathscr{G}\right)\log\pi\left(y_i\,\middle|\,x_i,\phi\right),\\Q_2^{\text{MF-like}}\left(\psi\,\middle|\,\theta^{(t)}\right)&=\sum_{i\in\mathscr{S}}\sum_{x_i}\pi\left(x_i\,\middle|\,\mathbf{X}_{\mathscr{N}(i)}=\tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)},y_i,\theta^{(t)},\mathscr{G}\right)\\&\qquad\log\pi\left(x_i\,\middle|\,\mathbf{X}_{\mathscr{N}(i)}=\tilde{\mathbf{x}}^{(t)}_{\mathscr{N}(i)},\psi,\mathscr{G}\right).\end{aligned}$$

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 neighbors $\mathbf{X}_{\mathscr{N}(i)}$ are fixed to their mean value, or more precisely $\tilde{\mathbf{x}}$ is set to the mean field estimate of the complete conditional distribution $\pi(\mathbf{x}\,|\,\mathbf{y},\theta,\mathscr{G})$, 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 $\pi(\cdot\,|\,\mathbf{y},\theta^{(t)},\mathscr{G})$ 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 $\pi(\cdot\,|\,\psi,\mathscr{G})$. 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

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**
Neighborhood restoration: draw $\tilde{\mathbf{x}}^{(t)}$ from $\pi\left(\cdot \,\middle|\, \mathbf{y}, \psi^{(t-1)}, \mathscr{G}\right)$;
E-step: compute

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

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

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

provides better results for the estimation of the spatial parameter, as illustrated in Forbes and Fort (2007).

1.7 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 2 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 1.4, within a Bayesian approach. However the

approximation produced has a variability significantly lower than the true posterior. Chapter 2 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).

Posterior parameter distribution

From a Bayesian perspective the focus is on the posterior parameter distribution. In Chapter 2, 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 $\pi(\mathbf{x} \mid \psi, \mathcal{G})$.

The posterior parameter distribution is

$$\pi\left(\psi \,\middle|\, \mathbf{x}^{\text{obs}}, \mathcal{G}\right) \propto \pi\left(\mathbf{x}^{\text{obs}} \,\middle|\, \psi, \mathcal{G}\right) \pi(\psi). \tag{1.30}$$

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

1.7.1 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 (1.30). 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

$$r\left(\psi' \,\middle|\, \psi\right) = \frac{\pi\left(\psi' \,\middle|\, \mathbf{x}, \mathscr{G}\right)}{\pi\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right)} \frac{\nu\left(\psi \,\middle|\, \psi'\right)}{\nu\left(\psi' \,\middle|\, \psi\right)} = \frac{Z\left(\psi, \mathscr{G}\right)}{Z\left(\psi', \mathscr{G}\right)} \frac{\pi(\psi') q\left(\mathbf{x} \,\middle|\, \psi', \mathscr{G}\right) \nu\left(\psi \,\middle|\, \psi'\right)}{\pi(\psi) q\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) \nu\left(\psi' \,\middle|\, \psi\right)}, \quad (1.31)$$

where $\nu(\theta \mid \theta')$ 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 (1.18), 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 (1.31) 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).

1.7.2 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 (1.31), we remark that the intractable ratio $\frac{Z(\psi, \mathscr{G})}{Z(\psi', \mathscr{G})}$ is replaced by $\frac{q(\mathbf{u}|\psi, \mathscr{G})}{q(\mathbf{u}|\psi', \mathscr{G})}$. 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 $\mathbf{U}$, namely $\pi(\cdot \mid \mathbf{x}, \psi)$, that makes it difficult to calibrate (see for

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 $\left\{\left(\psi^{(t)}, \psi'^{(t)}, \mathbf{u}^{(t)}\right)\right\}_{t=1}^{T}$

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.

1.8 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 neighborhood 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 1.8.2 brings a solution in the Bayesian paradigm which is explored in Chapter 4. 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 1.8.3 and discussed further in Chapter 4.

1.8.1 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

$$e(\mathbf{y} \mid m) = \int_{\Theta_m} \pi_m(\mathbf{y} \mid \theta_m)\pi_m(\theta_m)\mathrm{d}\theta_m. \tag{1.32}$$

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

$$\widehat{m}_{\text{MAP}}(\mathbf{y}^{\text{obs}}) = \arg\max_{m} \pi(m \mid \mathbf{y}^{\text{obs}}). \tag{1.33}$$

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

$$\pi_m(\mathbf{y} \mid \theta_m) = \int_{\mathscr{X}} \pi(\mathbf{y} \mid \mathbf{x}, \phi_m)\pi(\mathbf{x} \mid \psi_m, \mathscr{G}_m)\mu(\mathrm{d}\mathbf{x}), \tag{1.34}$$

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.

1.8.2 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 (1.30), namely

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

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 3.

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 neighborhood 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

Table 1.2: 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

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 1.2 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* $2^{100} \approx 10^{30}$*.*

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 (1.33): it approximates the posterior probabilities by the frequency of each model number associated with simulated $\mathbf{y}$'s in a neighborhood of $\mathbf{y}^{\text{obs}}$. If required, we can eventually predict the best model with the most frequent model in the neighborhood, or, in other words, take the final decision by plugging in (1.33) the approximations of the posterior probabilities.

At this stage, this first, naive algorithm faces the curse of dimensionality illustrated in Example 1.8.2. 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.

1.8.3 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 procedure 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 1.8.1, 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 (1.33) is equivalent to choose the model with the largest evidence (1.32). 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

$$\mathrm{BIC}(m) = -2\log \pi_m\left(\mathbf{y} \,\middle|\, \hat{\theta}_{\mathrm{MLE}}\right) + d_m \log(n) \approx -2\log \pi(\mathbf{y} \mid m), \tag{1.35}$$

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 4 for a more detailed presentation. The approximation may seem somewhat crude due to this $\mathscr{O}(1)$ 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 1.5.2) nor the incomplete likelihood (1.34) 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 1.5.2 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 (1.34), 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 1.5.2), namely

$$\pi\left(\mathbf{x} \,\middle|\, \psi_m, \mathscr{G}\right) \approx \mathbf{P}^{\text{MF-like}}\left(\mathbf{x} \,\middle|\, \psi_m, \mathscr{G}\right) = \prod_{i \in \mathscr{S}} \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}_{\mathscr{N}(i)}, \psi_m, \mathscr{G}\right). \tag{1.36}$$

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

$$\text{BIC}^{\text{MF-like}}(m) = -2 \log \mathbf{P}_m^{\text{MF-like}}\left(\mathbf{y} \,\middle|\, \hat{\theta}_{\text{MLE}}\right) + d_m \log(n). \tag{1.37}$$

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 1.5.2. As regards neighborhood 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 (1.34) 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 $\theta^{(t)}$

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

Afterwards the parameter is updated given the new configuration $\tilde{\mathbf{x}}^{(t+1)}$, the spatial component being updated by maximizing the pseudolihood (1.17),

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

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

$$\mathrm{PLIC}(m) = -2\log\left\{ \prod_{i\in\mathscr{S}} \sum_{\mathbf{x}_i} \pi\left(y_i \mid x_i, \phi_m^{\mathrm{ICM}}\right) \right. \\ \left. \pi\left(x_i \,\middle|\, \mathbf{X}_{\mathscr{N}(i)} = \mathbf{x}_{\mathscr{N}(i)}^{\mathrm{ICM}}, \psi_m^{\mathrm{ICM}}, \mathscr{G}\right) \right\} + d_m \log(n). \quad (1.38)$$

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 $Z(\psi, \mathscr{G})$ and $Z(\theta, \mathscr{G})$ denote the respective normalizing constants of the latent and the conditional fields (see Section 1.1.4), namely,

$$Z\left(\psi, \mathscr{G}\right) = \sum_{\mathbf{x}\in\mathscr{X}} \exp\left\{-H\left(\mathbf{x} \mid \psi, \mathscr{G}\right)\right\},$$
$$Z\left(\theta, \mathscr{G}\right) = \sum_{\mathbf{x}\in\mathscr{X}} \exp\left\{-H\left(\mathbf{x} \mid \mathbf{y}, \phi, \psi, \mathscr{G}\right)\right\} = \sum_{\mathbf{x}\in\mathscr{X}} \pi\left(\mathbf{y} \mid \mathbf{x}, \phi\right) \exp\left\{-H\left(\mathbf{x} \mid \psi, \mathscr{G}\right)\right\}.$$

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

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

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

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

The expression (1.35) turns into

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

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

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

where $Z^{\mathrm{MF}}(\psi, \mathscr{G})$ and $H^{\mathrm{MF}}(\mathbf{x} \mid \psi, \mathscr{G})$ 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

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

It follows from the positivity of the Kullback-Leibler divergence

$$Z\left(\psi, \mathscr{G}\right) \geq Z^{\mathrm{MF}}\left(\psi, \mathscr{G}\right) \exp\left(-\mathbf{E}^{\mathrm{MF}}\left\{H\left(\mathbf{X} \,\middle|\, \psi, \mathscr{G}\right) - H^{\mathrm{MF}}\left(\mathbf{X} \,\middle|\, \psi, \mathscr{G}\right)\right\}\right). \tag{1.39}$$

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 $\pi(\cdot \mid \mathbf{y}, \theta, \mathscr{G})$ 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

$$\begin{aligned}
\mathrm{BIC}^{\mathrm{GBF}}(m) = &-2\log\left\{Z^{\mathrm{MF}}\left(\hat{\theta}_m^{\mathrm{MLE}}, \mathscr{G}\right)\right\} + 2\log\left\{Z^{\mathrm{MF}}\left(\hat{\psi}_m^{\mathrm{MLE}}, \mathscr{G}\right)\right\} \\
&+ 2\mathbf{E}^{\mathrm{MF}}\left\{H\left(\mathbf{X} \,\middle|\, \mathbf{y}, \hat{\theta}_m^{\mathrm{MLE}}, \mathscr{G}\right) - H^{\mathrm{MF}}\left(\mathbf{X} \,\middle|\, \mathbf{y}, \hat{\theta}_m^{\mathrm{MLE}}, \mathscr{G}\right)\right\} \\
&- 2\mathbf{E}^{\mathrm{MF}}\left\{H\left(\mathbf{X} \,\middle|\, \hat{\psi}_m^{\mathrm{MLE}}, \mathscr{G}\right) - H^{\mathrm{MF}}\left(\mathbf{X} \,\middle|\, \hat{\psi}_m^{\mathrm{MLE}}, \mathscr{G}\right)\right\} \\
&+ d_m \log(n).
\end{aligned} \tag{1.40}$$

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.

2 Adjustment of posterior parameter distribution approximations

Computing the posterior distribution of model parameters (1.30) is a key objective in Bayesian analysis. Much of the literature has been devoted to Monte Carlo approaches that are time consuming (see Section 1.7). An alternative when dealing with complex model as Markov random fields is to use composite likelihoods, a class of non-genuine probability distributions that extend the pseudolikelihood (Besag, 1975), for Bayesian inference. This opportunity has been explored by Friel (2012) when dealing with the autologistic model. However the resulting approximation has a much lower variability than the true posterior distribution. Our main contribution is to present an approach to correct the posterior distribution resulting from using a misspecified likelihood function. The Chapter is organised as follows. Composite likelihoods are introduced in Section 2.1.1 before defining the composite posterior distribution in Section 2.1.2. In this Chapter we focus especially on how to formulate conditional composite likelihoods for application to the autologistic model. Our proposal to answer the issue of calibrating the composite likelihood function is developed through Section 2.2. Section 2.3 illustrates the performance of the various estimators for simulated data.

2.1 Bayesian inference using composite likelihoods

2.1.1 Composite likelihood

Composite likelihood methods are a natural extension of pseudolikelihood (Besag, 1975, see Section 1.4) that has encountered considerable interests in the statistics literature. The reader may refer to Varin et al. (2011) for a recent overview but we could mention headings such as pairwise likelihood methods (*e.g.*, Nott and Rydén, 1999), composite likelihood (*e.g.* Heagerty and Lele, 1998) and split-data likelihood

(*e.g.* Rydén, 1994) to name a few. Composite likelihoods are originally motivated by the fact that maximum pseudolikelihood estimator has generally larger asymptotic variance than maximum likelihood estimator and does not achieve the Cramer-Rao lower bound (Lindsay, 1988). Furthermore, empirical experiences have shown that the pseudolikelihood leads to unreliable estimates (*e.g.*, Geyer, 1991, Rydén and Titterington, 1998, Friel and Pettitt, 2004, Cucala et al., 2009).

Remember that $\mathscr{S} = \{1, \ldots, n\}$ is the index set for the graph nodes. The pseudolikelihood is based on the finest partition of $\mathscr{S}$ but one could use any subsets of the power set of $\mathscr{S}$. The starting idea is to extend the product (1.17) to a product of tractable joint probability distribution of a small number of variables, two in the example of pairwise likelihood. Given an integer $C \leq n$, we denote $\left\{A(i)\right\}_{i=1}^{C} \subseteq \mathscr{P}(\mathscr{S})$ and $\left\{B(i)\right\}_{i=1}^{C} \subseteq \mathscr{P}(\mathscr{S})$ sets of subset of $\mathscr{S}$, that is each $A(i)$ or $B(i)$ corresponds to an index subset for the graph nodes. Following Asuncion et al. (2010), in its simple version a composite likelihood can be written as follows

$$f_{\mathrm{CL}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \prod_{i=1}^{C} \pi\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{B(i)}, \psi, \mathscr{G}\right). \tag{2.1}$$

It is worth mentioning some special cases.

(i) The trivial situation where the set $A(i)$ contains all the sites, that is $C = 1$, $A(i) = \mathscr{S}$, $B(i) = \emptyset$, , corresponds to the full likelihood.

(ii) When the product expresses only in terms of marginal distribution of the subset $A(i)$, that is $B(i) = \emptyset$, one usually talks about *marginal composite likelihood*.

(iii) When one takes for the subset $B(i)$ the absolute complement of $A(i)$, namely $B(i) = \mathscr{S} \setminus A(i)$, the function is often termed *conditional composite likelihood*. The pseudolikelihood is a particular case where the product is taken over all graph nodes, *i.e.*, $C = n$. Each $A(i)$ is then a singleton, namely $A(i) = \{i\}$, and $B(i)$ is the set of neighboring sites $\mathscr{N}(i)$.

In this Chapter, following Lindsay (1988), we suggest that a composite likelihood should take the general form

$$f_{\mathrm{CL}}^{\mathrm{cal}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \prod_{i=1}^{C} \pi\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{B(i)}, \psi, \mathscr{G}\right)^{w_i}, \tag{2.2}$$

where w_i are positive weights that are to specify in the present work.

Composite likelihood has been made popular in a context where marginal distributions can be computed (*e.g.*, Varin et al., 2011). Spatial lattice processes such as the autologistic model differ from that class of models in the sense that dependence structure makes impossible the calculation of marginal probabilities. Indeed factors thus involve integrals over the absolute complement of $A(i)$ that are of the same complexity than the cumbersome normalizing constant. Throughout the Chapter the focus is thus solely on conditional composite likelihood. We limit our study to models defined on a regular rectangular lattice of size $h \times w$, where h stands for the height and w for the width of the lattice, and whose Hamiltonian is of the form $H(\mathbf{x} \mid \psi, \mathscr{G}) = \psi^T \mathbf{S}(\mathbf{x})$ where $\mathbf{S}(\mathbf{x}) = \{s_1(\mathbf{x}), \ldots, s_d(\mathbf{x})\}$ is a vector of sufficient statistics. We shall remark that this vector of sufficient statistics is a function depending on $\mathscr{G}$ (see Example 2.1.1) even if it is omitted in the notation for the sake of simplicity. We restrict each $A(i)$ to be of the same dimension and in particular to correspond to contiguous square blocks of lattice points of size $k \times k$. In terms of the value of C in case *(iii)*, an exhaustive set of blocks would result in $C = (h-k+1) \times (w-k+1)$. In particular, we allow the collection of blocks $\{A(i)\}_{i=1}^{C}$ to overlap with one another.

The conditional composite likelihood relies on evaluating

$$\pi\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}, \psi, \mathscr{G}\right) = \frac{\exp\left(\psi^T \mathbf{S}\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right)\right)}{Z\left(\psi, \mathscr{G}, \mathbf{x}_{-A(i)}\right)},$$

where $\mathbf{S}\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right) = \left\{s_1\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right), \ldots, s_d\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right)\right\}$ is the restriction of $\mathbf{S}(\mathbf{x})$ to the subgraph defined on the set $A(i)$ and conditioned on the realised $\mathbf{x}_{-A(i)}$, that is conditioned by all the lattice point of $\mathbf{x}_{-A(i)}$ connected to a lattice point of $\mathbf{x}_{A(i)}$ by an edge of $\mathscr{G}$. This can be understood in terms of induced graph.

Definition 4. *The graph induced by $\mathscr{G}$ on the set of graph nodes $A(i)$, denoted $\Gamma(\mathscr{G}, A(i))$, is the undirected graph whose set of edges gathers the edges of $\mathscr{G}$ attached to at least one nodes in $A(i)$, i.e.,*

$$\ell \overset{\Gamma(\mathscr{G}, A(i))}{\sim} j \quad \Longleftrightarrow \quad \ell \overset{\mathscr{G}}{\sim} j \text{ with } \ell \in A(i) \text{ or } j \in A(i).$$

Hence, $\mathbf{S}\left(\cdot \,\middle|\, \mathbf{x}_{-A(i)}\right)$ is the function defined on the graph $\Gamma(\mathscr{G}, A(i))$ for which the value of nodes ℓ is set to x_ℓ if ℓ is not in $A(i)$.

Besides the normalizing constant now includes the argument $\mathbf{x}_{-A(i)}$ emphasising that it involves a summation over all possible realisations of sub-lattices defined on the set $A(i)$ conditionally to the realised $\mathbf{x}_{-A(i)}$, namely

$$Z\left(\psi, \mathscr{G}, \mathbf{x}_{-A(i)}\right) = \sum_{\tilde{\mathbf{x}}_{A(i)}} \exp\left(\psi^T \mathbf{S}\left(\tilde{\mathbf{x}}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right)\right).$$

Generalised recursions for computing the normalizing constant of general factorisable models such as the autologistic models (see Section 1.3) extends easily to allow on to compute the latter.

Example. *The autologistic model (see Section 1.1.2) is defined in terms of two sufficient statistics,*

$$s_0(\mathbf{x}) = \sum_{i=1}^{n} x_i, \text{ and } s_1(\mathbf{x}) = \sum_{i \overset{\mathscr{G}}{\sim} j} x_i x_j,$$

where $i \overset{\mathscr{G}}{\sim} j$ means that lattice points i and j are connected by an edge in $\mathscr{G}$. Henceforth we assume that the lattice points have been indexed from top to bottom in each column and where columns are ordered from left to right. For example, for a first order neighbourhood model an interior point x_i has neighbours $\{x_{i-h}, x_{i-1}, x_{i+1}, x_{i+h}\}$. Along the edges of the lattice each point has either 2 or 3 neighbours. The full-conditional of an inner lattice point x_i can be written as

$$\pi\left(x_i \,\middle|\, \mathbf{x}_{-i}, \psi, \mathscr{G}\right) \propto \exp\left(\alpha y_i + \beta\{x_i x_{i-h} + x_i x_{i-1} + x_i x_{i+1} + x_i x_{i+h}\}\right).$$

When considering conditional composite likelihood with blocks, the full-conditional distribution of $A(i)$ can be written as

$$\pi\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}, \psi, \mathscr{G}\right) = \frac{\exp\left(\alpha s_0\left(\mathbf{x}_{A(i)}\right) + \beta s_1\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right)\right)}{Z\left(\psi, \mathscr{G}, \mathbf{x}_{-A(i)}\right)},$$

where

$$s_0\left(\mathbf{x}_{A(i)}\right) = \sum_{j \in A(i)} x_j, \text{ and}$$
$$s_1\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right) = \sum_{\ell \overset{\Gamma(\mathscr{G}, A(i))}{\sim} j} x_\ell x_j = \sum_{\ell \overset{\mathscr{G}}{\sim} j} x_\ell x_j \left(\mathbf{1}\{\ell \in A(i)\} + \mathbf{1}\{j \in A(i)\}\right).$$

In the normalizing constant, we shall note that the expression of s_1 is slightly different

$$s_1\left(\tilde{\mathbf{x}}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right) = \sum_{\ell \overset{\mathscr{G}}{\sim} j} \left(\tilde{x}_\ell \mathbf{1}\{\ell \in A(i)\} + x_\ell \mathbf{1}\{\ell \notin A(i)\}\right) \left(\tilde{x}_j \mathbf{1}\{j \in A(i)\} + x_j \mathbf{1}\{j \notin A(i)\}\right).$$

The auto-models of Besag (1974) allow variations on the level of dependencies between edges and a potential anisotropy can be introduced on the graph. Indeed, consider a set of graphs $\{\mathscr{G}_1, \ldots, \mathscr{G}_M\}$ with a single parameter on the edges. Each graph of dependency

$\mathscr{G}_m$ induces a summary statistic

$$s_m(\mathbf{x}) = \sum_{j=1}^{n} \sum_{i \overset{\mathscr{G}_m}{\sim} j} x_i x_j .$$

For instance, one can consider an anisotropic configuration of a first order neighbourhood model: that is edges of $\mathscr{G}_1$ are all the vertical edges of the lattice and edges of $\mathscr{G}_2$ are all the horizontal ones. Then an interior point x_i has neighbours $\{x_{i-1}, x_{i+1}\}$ according to $\mathscr{G}_1$ and $\{x_{i-m}, x_{i+m}\}$ according to $\mathscr{G}_2$. This allows to set an interaction strength that differs according to the direction (see Table 1.1).

2.1.2 Conditional composite posterior distribution

Turning back to the issue of computing the posterior distribution (1.30), our proposal is to replace the true likelihood $\pi(\cdot \mid \psi, \mathscr{G})$ with a conditional composite likelihood, leading us to concentrate on the approximated posterior distribution, referred to as (calibrated) composite posterior distribution,

$$\pi_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right) \propto f_{\mathrm{CL}}^{\mathrm{cal}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) \pi\left(\psi\right). \tag{2.3}$$

For the sake of clarity, in the special case of $w_i = 1$, we refer to the above as non-calibrated composite posterior distribution and we denote it

$$\pi_{\mathrm{CL}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right) \propto f_{\mathrm{CL}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) \pi\left(\psi\right). \tag{2.4}$$

Currently, there is very little literature on the use of composite likelihoods in the Bayesian setting, although Pauli et al. (2011) and Ribatet et al. (2012) present a discussion on the use of marginal composite likelihoods in the Bayesian setting. From the standpoint of conditional composite likelihood, mention the work of Friel (2012) subsequent to a study conducted by Okabayashi et al. (2011) although from a maximum likelihood perspective. Friel (2012) is interested in the formulation of posterior approximations for the autologistic model. His proposal is to work with a product of conditional distribution of blocks of the lattice for which the normalizing constant can be computed using the recursion of Section 1.3. A similar approach has also been lead by Friel et al. (2009) in the context of hidden Markov random field.

In his related work, Friel (2012) has examined composite likelihood for various block sizes but only for $w_i = 1$. Whilst the approximation is easy to compute, we highlight here the empirical observation that non-calibrated composite likelihood leads to a composite posterior distribution with substantially lower variability than the true

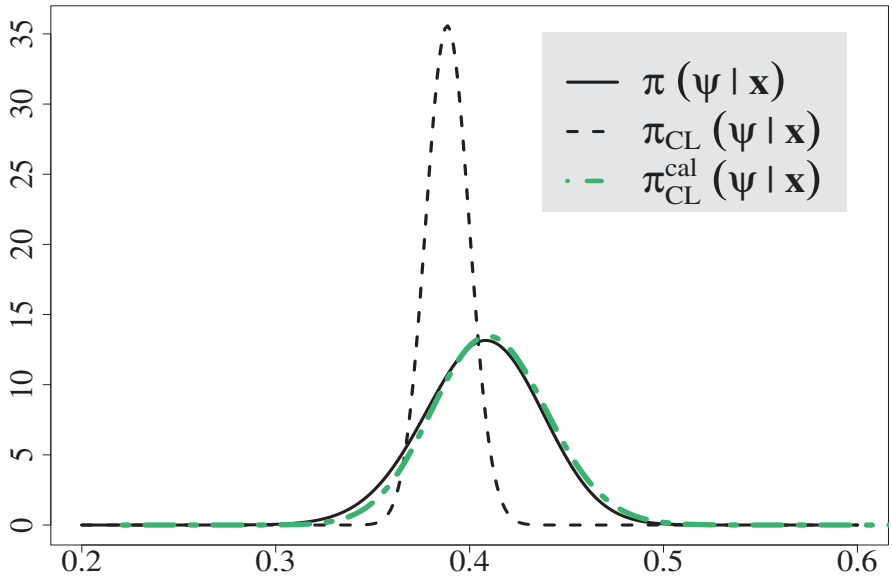


Figure 2.1: Posterior parameter distribution (plain), non-calibrated composite posterior distribution (dashed) and composite posterior distribution (green) with a uniform prior for a realization of the Ising model on a 16×16 lattice near the phase transition. The conditional composite likelihood is computed for an exhaustive set of 4×4 blocks.

posterior distribution, leading to overly precise posterior parameters, see Figure 2.1. The main contribution of this Chapter is to present an approach to tune the weights w_i in order to correct the posterior distribution resulting from using a misspecified likelihood function. In our context, there exists no particular reason to weight each block differently. Consequently we assume that each block has the same weight and we denote it w, so that

$$\pi_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \,|\, \mathbf{x}, \mathscr{G}\right) \propto \left\{ \prod_{i=1}^{C} \pi\left(\mathbf{x}_{A(i)} \,|\, \mathbf{x}_{B(i)}, \psi, \mathscr{G}\right) \right\}^{w} \pi\left(\psi\right).$$

2.1.3 Estimation algorithm of the Maximum *a posteriori*

Our proposal to adjust the non-calibrated composite posterior distribution relies on matching conditions about the posterior mode so that the estimation of maximum *a posteriori* is paramount to the approach. For this purpose, we propose to use a stochastic gradient algorithm (see also Section 1.5.1) to address the issue of estimating

$$\hat{\psi}_{\mathrm{MAP}} = \arg\max_{\psi} \pi\left(\psi \,|\, \mathbf{x}, \mathscr{G}\right) \quad \text{and} \quad \hat{\psi}_{\mathrm{CL}} = \arg\max_{\psi} \pi_{\mathrm{CL}}\left(\psi \,|\, \mathbf{x}, \mathscr{G}\right).$$

Using (1.19) and (1.20) as a starting point, we can write the gradient of the log-posterior for ψ as

$$\nabla \log \pi\left(\psi \mid \mathbf{x}, \mathscr{G}\right) = \mathbf{S}(\mathbf{x}) - \mathbf{E}_{\psi}\left\{\mathbf{S}\left(\mathbf{X}\right)\right\} + \nabla \log \pi(\psi), \tag{2.5}$$

where $\mathbf{E}_{\psi}\{\mathbf{S}(\mathbf{X})\}$ denotes the expected value of $\mathbf{S}(\mathbf{X})$ with respect to $\pi(\cdot \mid \psi, \mathscr{G})$, and the Hessian matrix of the log-posterior for ψ,

$$\nabla^2 \log \pi\left(\psi \mid \mathbf{x}, \mathscr{G}\right) = -\mathbf{Var}_{\psi}\left\{\mathbf{S}(\mathbf{X})\right\} + \nabla^2 \log \pi(\psi), \tag{2.6}$$

where $\mathbf{Var}_{\psi}\{\mathbf{S}(\mathbf{X})\}$ denotes the covariance matrix of $\mathbf{S}(\mathbf{X})$ with respect to $\pi(\cdot \mid \psi, \mathscr{G})$. Addressing the issue of estimation of $\hat{\psi}_{\text{MAP}}$, we note generally from equation (2.6) that $\log \pi(\cdot \mid \mathbf{x}, \mathscr{G})$ is not a concave function and one can ask if the assumption of a single mode is valid. The Hessian of the log-likelihood is a semi-negative matrix for any ψ, namely $-\mathbf{Var}_{\psi}\{\mathbf{S}(\mathbf{X})\}$, and so the log-likelihood is uni-modal. A reasonable choice of prior, for example with $\nabla^2 \log \pi(\psi)$ a semi-negative Hessian matrix for any ψ, will thus lead to a uni-modal, or at least locally concave, posterior distribution.

Similar to (2.5) and (2.6), one can express the gradient and Hessian of the composite log-posterior $\log \pi_{\text{CL}}(\cdot \mid \mathbf{x}, \mathscr{G})$ in terms of moments of statistics, namely

$$\nabla \log \pi_{\text{CL}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right) = \sum_{i=1}^{C} \mathbf{S}\left(\mathbf{x}_{A(i)} \mid \mathbf{x}_{-A(i)}\right) - \mathbf{E}_{\psi}\left\{\mathbf{S}\left(\mathbf{X}_{A(i)} \mid \mathbf{x}_{-A(i)}\right)\right\} + \nabla \log \pi(\psi),$$

$$\nabla^2 \log \pi_{\text{CL}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right) = -\sum_{i=1}^{C} \mathbf{Var}_{\psi}\left\{\mathbf{S}\left(\mathbf{X}_{A(i)} \mid \mathbf{x}_{-A(i)}\right)\right\} + \nabla^2 \log \pi(\psi).$$

Hence conclusion towards unicity of the maximum and uni-modality of the distribution remains the same for the composite posterior distribution $\pi_{\text{CL}}(\cdot \mid \mathbf{x}, \mathscr{G})$.

Equation (2.5) suggests that it is possible to estimate the gradient of the posterior for ψ using Monte Carlo draws from $\pi(\cdot \mid \psi, \mathscr{G})$ (see Section 1.2). On this basis, in the frequentist paradigm, Younes (1988) sets a stochastic gradient algorithm to estimate the maximum likelihood of Markov random fields with finite number of states. At each step t of the algorithm the expectation $\mathbf{E}_{\psi^{(t)}}\{\mathbf{S}(\mathbf{X})\}$ is replaced by the value of the statistic function for a realization of the random field $\mathbf{x}^{(t)}$. The update of parameters at iteration t of the gradient descent can thus be written as

$$\psi^{(t+1)} = \psi^{(t)} + \frac{\delta}{t+1}\left\{\mathbf{S}(\mathbf{x}) - \mathbf{S}\left(\mathbf{x}^{(t)}\right) + \nabla \log \pi\left(\psi^{(t)}\right)\right\},$$

where the step size depends on a threshold δ Younes (1988) demonstrated the convergence of the algorithm for small enough δ. In practical terms the theoretical value of

δ yields a step size too small to ensure the convergence to be achieved in reasonable amount of time. Instead following a remark of Younes (1988), we consider a step of the form $\frac{1}{t+n_0}$. The integer n_0 controls the probability of non-convergence and shall be chosen large enough. Indeed care must be taken with the optimization algorithm especially when applied to the composite posterior distribution since it is typically very sharp around the mode, as shown in Figure 2.1. Hence for a too large step size, algorithm oscillates a long time before converging. To avoid this phenomenon one usually sets for example n_0 to 1000. Nevertheless using solely one realization per iteration is somewhat crude especially if the statistic $\mathbf{S}\left(\mathbf{X}^{(t)}\right)$ has a great variability. We advocate in favour of the gradient descent based on a Monte-Carlo estimator of the expectation, namely

$$\psi^{(t+1)} = \psi^{(t)} + \frac{1}{t+n_0}\left\{\mathbf{S}(\mathbf{x}) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{S}\left(\mathbf{x}_j^{(t)}\right) + \nabla \log \pi\left(\psi^{(t)}\right)\right\},$$

where $\mathbf{x}_j^{(t)}$ are drawn from $\pi(\cdot \mid \psi^{(t)}, \mathscr{G})$.

In our experiments we have found that using a Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm which is based on a Hessian matrix approximation using rank-one updates calculated from approximate gradient evaluations, can provide good performance but lack theoretical support. The scheme can be described as

$$\psi^{(t+1)} = \psi^{(t)} + \delta^{(t)} B^{(t)}\left\{\mathbf{S}(\mathbf{x}) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{S}\left(\mathbf{x}_j^{(t)}\right) + \nabla \log \pi\left(\psi^{(t)}\right)\right\},$$

where $\delta^{(t)}$ is the step size and $B^{(t)}$ is a matrix approximating the Hessian. Adding the information contained in the Hessian can speed up the convergence of the algorithm but requires a careful choice of the step size $\delta^{(t)}$ otherwise the algorithm might be numerically unstable. One generally uses golden section search or line search in the direction

$$B^{(t)}\left\{\mathbf{S}(\mathbf{x}) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{S}\left(\mathbf{x}_j^{(t)}\right) + \nabla \log \pi\left(\psi^{(t)}\right)\right\}$$

to find an acceptable step size. This solution is obviously not suitable to our context since it requires to evaluate several times the function and the gradient of the function to maximize. Our strategy is to use the information contained in the prior $\pi(\psi)$ to define a convex compact set $\mathscr{D}$, namely a subset of the parameter space such that $\mathbf{P}(\psi \in \mathscr{D}) \geq \alpha$, with α in $]0,1]$. The step size is then design so that updates remain in $\mathscr{D}$,

that is given a norm $\|\cdot\|$ on $\mathbb{R}^d$ we choose $\delta^{(t)}$ such that

$$\left\| \delta^{(t)} B^{(t)} \left\{ \mathbf{S}(\mathbf{x}) - \frac{1}{N} \sum_{j=1}^{N} \mathbf{S}\left(\mathbf{x}_j^{(t)}\right) + \nabla \log \pi\left(\psi^{(t)}\right) \right\} \right\| \leq \max\{\|a-b\|,\ (a,b) \in \mathscr{D}\}.$$

The primary drawback of this stochastic gradient algorithm lies in the Monte Carlo draws. When applying this algorithm to composite posterior distribution, it is possible to use recursions of Section 1.3 as outlined in Friel and Rue (2007) to draw exactly from $\pi(\mathbf{x}_{A(i)} \mid \mathbf{x}_{-A(i)}, \psi, \mathscr{G})$ instead. Finally estimating $\hat{\psi}_{\text{MAP}}$ using a random initialization point in gradient algorithm is inefficient. Indeed, estimating $\mathbf{E}_{\psi}\{\mathbf{S}(\mathbf{X})\}$ is the most cumbersome part of the algorithm since it involves sampler such as the Gibbs sampler or the Swendsen-Wang algorithm and should be done as little as possible. Despite that $\hat{\psi}_{\text{CL}}$ differs from $\hat{\psi}_{\text{MAP}}$ it is usually close and turns out to yield a good initialization to the optimization algorithm.

2.1.4 On the asymptotic theory for composite likelihood inference

The calibration question of the composite likelihood has long-standing antecedents in the frequentist paradigm (Varin et al., 2011, Section 2.3 and the references therein). In this context, the adjustments are required to recover the asymptotic chi-squared distribution of the likelihood ratio statistic. To precise this statement, consider r independent and identically distributed observations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(r)}$ from the statistical model $\pi(\cdot \mid \psi, \mathscr{G})$ and denote $d = \dim(\Psi)$ the dimension of the parameter space. Under regularity conditions, it follows by a simple Taylor series expansion that

$$W(\psi) = 2 \sum_{j=1}^{r} \left\{ \log \pi\left(\mathbf{x}^{(j)} \,\middle|\, \hat{\psi}_{\text{MLE}}, \mathscr{G}\right) - \log \pi\left(\mathbf{x}^{(j)} \,\middle|\, \psi, \mathscr{G}\right) \right\} \underset{r \to +\infty}{\longrightarrow} \chi^2_d,$$

where $\hat{\psi}_{\text{MLE}} = \arg\max_{\psi} \sum_{j=1}^{r} \log \pi(\mathbf{x}^{(j)} \mid \psi, \mathscr{G})$ is the maximum likelihood estimator (Wilks, 1938). This result is slightly modified for misspecified models, that is when the likelihood is replaced with non-calibrated composite likelihoods (2.1) in the above. Within that framework, the score function is replaced with the composite score function which is the linear combination of the scores associated to each marginal or conditional densities of (2.1)

$$\nabla \log f_{\text{CL}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) = \sum_{i=1}^{C} \nabla \log \pi\left(\mathbf{x}_{A(i)} \,\middle|\, \mathbf{x}_{B(i)}, \psi, \mathscr{G}\right).$$

With respect to this score function, denote,

- $\hat{\psi}_{\text{MCLE}}$ the maximum composite likelihood estimator, that is the solution of the composite score function $\sum_{j=1}^{r} \nabla \log f_{\text{CL}}(\mathbf{x}^{(j)} \mid \psi, \mathscr{G})$,

- $H(\psi)$ the sensitivity matrix defined as the Hessian of the composite log-likelihood $\log f_{\text{CL}}^{\text{cal}}(\cdot \mid \psi, \mathscr{G})$ with respect to the Gibbs distribution $\pi(\cdot \mid \psi, \mathscr{G})$:

$$\begin{aligned} H(\psi) &= \mathbf{E}_{\psi}\left\{-\nabla^2 \log f_{\text{CL}}\left(\mathbf{X} \,\middle|\, \psi, \mathscr{G}\right)\right\} \\ &= -\int_{\mathscr{X}} \nabla^2 \log f_{\text{CL}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) \pi\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right) \mu(\mathrm{d}\mathbf{x}), \end{aligned}$$

- $J(\psi)$ the variability matrix defined as the covariance matrix of the composite score function with respect to the Gibbs distribution $\pi(\cdot \mid \psi, \mathscr{G})$:

$$J(\psi) = \mathbf{Var}_{\psi}\left\{\nabla \log f_{\text{CL}}\left(\mathbf{X} \,\middle|\, \psi, \mathscr{G}\right)\right\}.$$

The asymptotic distribution of the composite likelihood ratio statistic is a linear combination of independent chi-squared variates $Z_1, \ldots, Z_d$ (*e.g.*, Varin et al., 2011), namely

$$2\sum_{j=1}^{r}\left\{\log f_{\text{CL}}\left(\mathbf{x}^{(j)} \,\middle|\, \hat{\psi}_{\text{MCLE}}, \mathscr{G}\right) - \log f_{\text{CL}}\left(\mathbf{x}^{(j)} \,\middle|\, \psi, \mathscr{G}\right)\right\} \underset{r\to+\infty}{\longrightarrow} \sum_{j=1}^{d} \lambda_j(\psi) Z_j,$$

where $\lambda_1(\psi), \ldots, \lambda_d(\psi)$ are the eigenvalues of $H(\psi)^{-1}J(\psi)$. The non-standard asymptotic null distribution is due to that the maximum composite likelihood estimator is consistant but has an asymptotic variance larger than the maximum likelihood estimator. Indeed, under regularity conditions, the maximum composite likelihood estimator is asymptocally normally distributed,

$$\sqrt{r}\left(\hat{\psi}_{\text{MCLE}} - \psi\right) \underset{r\to+\infty}{\longrightarrow} \mathscr{N}_d\left(0, H^{-1}(\psi) J(\psi) H(\psi)^{-1}\right),$$

where $\mathscr{N}_d(\cdot,\cdot)$ denotes the d-dimensional normal distribution (*e.g.*, Kent, 1982, Lindsay, 1988). In the context of single time series or random field, it may be interesting to have asymptotic results when the number of replicates is fixed (usually $r = 1$) and the observation size n grows to infinity (*e.g.*, Geman and Graffigne, 1986, Cox and Reid, 2004). The asymptotic properties depend on ergodicity conditions and r is replaced by n in the above.

Pauli et al. (2011) and Ribatet et al. (2012) independently suggest to use adjusted composite likelihood functions to define a composite posterior distribution (2.3) and establish the asymptotic normality of the latter. The corrections used are of two kinds. The first is a moment matching solution (Geys et al., 2001) which ensures

that the expectation of the adjusted composite likelihood ratio converges toward the expectation of the chi-squared distribution. The second is a scaling solution (Chandler and Bate, 2007) that recovers the right asymptotic null distribution by modifying the curvature of the composite likelihood. Both modifications rely on the evaluation of the sensitivity matrix $H(\psi)$ and the variability matrix $J(\psi)$ but do not include any information about the true likelihood or the prior. In our context, a calibration based on asymptotic efficiency is questionable since it is possible to get punctual estimates of the gradient and the Hessian of the log-posterior distribution on the basis of equations (1.19) and (1.20).

2.2 Conditional composite likelihood adjustments

Our proposal is to make a shift from asymptotical behaviour to local matching conditions to calibrate the weight w. The following Sections provide modifications that aim at correcting the mode and the variance of a sample drawn from a composite posterior distribution by adjusting the mode and the curvature at the mode of the latter. The idea behind has appeared in other contexts such as Gaussian Markov random fields (*e.g.*, Rue et al., 2009).

2.2.1 Magnitude adjustment

The general approach we propose to adjust the covariance of the composite posterior is to temper the conditional composite likelihood with some weight w in order to modify its curvature around the mode. We remark that the curvature of a scalar field at its maximum is directly linked to the Hessian matrix. Based on that observation, our proposal is to choose w such that

$$\nabla^2 \log \pi\left(\hat{\psi}_{\text{MAP}} \,\middle|\, \mathbf{x}, \mathscr{G}\right) = w \nabla^2 \log \pi_{\text{CL}}\left(\hat{\psi}_{\text{CL}} \,\middle|\, \mathbf{x}, \mathscr{G}\right), \tag{2.7}$$

where both $\hat{\psi}_{\text{MAP}}$ and $\hat{\psi}_{\text{CL}}$ are computed with the stochastic gradient algorithm of Section 2.1.3.

For a homogeneous and isotropic Markov random field without potential on singletons such as the Ising model, the model parameter is scalar ($\psi \equiv \beta \in \mathbb{R}$) which yields a simple expression for w, namely

$$w = \frac{\mathbf{Var}_{\hat{\psi}_{\text{MAP}}}\{\mathbf{S}(\mathbf{X})\} - \nabla^2 \log \pi\left(\hat{\psi}_{\text{MAP}}\right)}{\sum_{i=1}^{C} \mathbf{Var}_{\hat{\psi}_{\text{CL}}}\left\{\mathbf{S}\left(\mathbf{X}_{A(i)} \,\middle|\, \mathbf{x}_{-A(i)}\right)\right\} - \nabla^2 \log \pi\left(\hat{\psi}_{\text{CL}}\right)}. \tag{2.8}$$

Table 2.1: Weight options for a magnitude adjustment in presence of anisotropy or potential on singletons ($\psi \in \mathbb{R}^d$)

Weight	Definition
$w^{(1)}$	$\frac{\text{tr}\{\nabla^2 \log \pi(\hat{\psi}_{\text{MAP}} \mid \mathbf{x}, \mathscr{G})\}}{\text{tr}\{\nabla^2 \log \pi_{\text{CL}}(\hat{\psi}_{\text{CL}} \mid \mathbf{x}, \mathscr{G})\}}$
$w^{(2)}$	$\frac{1}{d}\sum_{j=1}^{d} \frac{\mathbf{Var}_{\hat{\psi}_{\text{MAP}}}\{s_j(\mathbf{X})\} - \{\nabla^2 \log \pi(\hat{\psi}_{\text{MAP}})\}_{jj}}{\sum_{i=1}^{C} \mathbf{Var}_{\hat{\psi}_{\text{CL}}}\{s_j(\mathbf{X}_{A(i)} \mid \mathbf{x}_{-A(i)})\} - \{\nabla^2 \log \pi(\hat{\psi}_{\text{CL}})\}_{jj}}$
$w^{(3)}$	$\frac{1}{d}\text{tr}\left[\nabla^2 \log \pi\left(\hat{\psi}_{\text{MAP}} \mid \mathbf{x}, \mathscr{G}\right)\left\{\nabla^2 \log \pi_{\text{CL}}\left(\hat{\psi}_{\text{CL}} \mid \mathbf{x}, \mathscr{G}\right)\right\}^{-1}\right]$
$w^{(4)}$	$\left[\frac{\det\{\nabla^2 \log \pi(\hat{\psi}_{\text{MAP}} \mid \mathbf{x}, \mathscr{G})\}}{\det\{\nabla^2 \log \pi_{\text{CL}}(\hat{\psi}_{\text{CL}} \mid \mathbf{x}, \mathscr{G})\}}\right]^{\frac{1}{d}}$
$w^{(5)}$	$\frac{\lVert \nabla^2 \log \pi(\hat{\psi}_{\text{MAP}} \mid \mathbf{x}, \mathscr{G}) \rVert_F}{\lVert \nabla^2 \log \pi_{\text{CL}}(\hat{\psi}_{\text{CL}} \mid \mathbf{x}, \mathscr{G}) \rVert_F}$, where $\lVert \cdot \rVert_F$ is the Frobenius norm

However in presence of anisotropy or potential on singletons, the system of equations 2.7 is overdetermined such that a unique solution does not exist in general. Thus, we have explored different necessary conditions for fulfilling the scalar constraint between the two Hessian matrices (see Table 2.1). The weights $w^{(1)}$ and $w^{(2)}$ neglect the covariance between the summary statistics by including only the information contained in the diagonal of each matrix whereas weights $w^{(3)}$, $w^{(4)}$ and $w^{(5)}$ take advantage of these covariances.

2.2.2 Curvature adjustment

The magnitude adjustment aims at scaling the composite posterior distribution down to the appropriate magnitude by performing a non-linear transformation of vertical axis. The weight w similarly affects each direction of space parameters leaving the overall geometry unchanged. The latter does not take into account a possible modification of the correlation between the variables induced by the use of a composite likelihood. We expect this phenomenon to be particularly important when dealing with models where there is a potential on singletons such as the autologistic model. Indeed estimations of the abundance parameter α and interaction parameter β do not suffer from the same level of approximation relating to the independence assumption between blocks. Thus we should move from the general form (2.2) with a scalar weight on blocks to one involving a matrix of weights.

We follow a suggestion of Chandler and Bate (2007) who, in the context of hypothesis testing, modify the curvature of the composite likelihood around its global maximum by

$$f_{\mathrm{CL}}^{\mathrm{cal}}\left(\mathbf{x} \mid \psi, \mathscr{G}\right) = f_{\mathrm{CL}}\left(\mathbf{x} \mid \hat{\psi}_{\mathrm{MCLE}} + W(\psi - \hat{\psi}_{\mathrm{MCLE}})\right),$$

for some constant $d \times d$ matrix W. While the substitution keeps the same maximum, it deforms the geometry of the parameter space through the matrix W by stretching linearly the horizontal axis. Hereafter, the resulting composite posterior is referred to as

$$\pi_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}, W\right) \propto f_{\mathrm{CL}}\left(\mathbf{x} \mid \hat{\psi}_{\mathrm{CL}} + W(\psi - \hat{\psi}_{\mathrm{CL}})\right) \pi\left(\psi\right).$$

Chandler and Bate (2007) set up the matrix W such that the composite likelihood ratio has an asymptotic chi-squared distribution, which leads to choose W such that

$$W^T H\left(\hat{\psi}_{\mathrm{MCLE}}\right) W = H\left(\hat{\psi}_{\mathrm{MCLE}}\right) J\left(\hat{\psi}_{\mathrm{MCLE}}\right)^{-1} H\left(\hat{\psi}_{\mathrm{MCLE}}\right).$$

We rather focus on the covariance matrix at the estimated maximum *a posteriori*. Indeed, we follow the approach introduced in Section 2.2.1 and we choose W such that,

$$\nabla^2 \log \pi\left(\psi \mid \mathbf{x}, \mathscr{G}\right)\big|_{\psi=\hat{\psi}_{\mathrm{MAP}}} = \nabla^2 \log \pi_{\mathrm{CL}}\left(\hat{\psi}_{\mathrm{CL}} + W(\psi - \hat{\psi}_{\mathrm{CL}})\right)\big|_{\psi=\hat{\psi}_{\mathrm{CL}}}$$

which is equivalent to

$$\nabla^2 \log \pi\left(\hat{\psi}_{\mathrm{MAP}} \mid \mathbf{x}, \mathscr{G}\right) = W^T \nabla^2 \log \pi_{\mathrm{CL}}\left(\hat{\psi}_{\mathrm{CL}}\right) W. \tag{2.9}$$

The choice of W is not unique due to the absence of uniqueness of the square root of a matrix. This problem is also encountered by Ribatet et al. (2012) who suggest to take any semi-definite negative matrix. In what follows, we assume that W is a lower triangular matrix. In practice we have observe that any lower triangular matrix solution yield almost equivalent performances. We would like to draw reader's attention to the fact that the choice of W deserves to be further probed.

2.2.3 Mode adjustment

Once the composite posterior distribution adjusted, it remains the issue of what we could call a mode misspecification. A drawback with composite posterior distribution in most cases is the bias with the maximum *a posteriori* $\hat{\psi}_{\mathrm{MAP}}$. In this Section, $\pi_{\mathrm{CL}}^{\mathrm{cal}}$

stands for any of the composite posterior distribution resulting from a magnitude or curvature adjustment unless otherwise specified. The present modification ensures that the posterior and the composite posterior have the same mode. Under reasonable assumptions on the prior, these distributions have a unique maximum and the adjustment is simply the substitution

$$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right) = \pi_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi - \hat{\psi}_{\mathrm{MAP}} + \hat{\psi}_{\mathrm{cal}} \,\middle|\, \mathbf{x}, \mathscr{G}\right), \tag{2.10}$$

$\hat{\psi}_{\mathrm{cal}}$ is the maximum *a posteriori* of the adjusted composite posterior distribution, namely

$$\hat{\psi}_{\mathrm{cal}} = \underset{\psi}{\arg\max}\ \pi_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right).$$

The value of $\hat{\psi}_{\mathrm{MAP}}$ being estimated upstream the computation of w, we solely have to compute $\hat{\psi}_{\mathrm{cal}}$. This is done for a low computational cost by once again applying the stochatic gradient algorithm of Section 2.1.3 to $\log \pi_{\mathrm{CL}}^{\mathrm{cal}}(\cdot \mid \mathbf{x}, \mathscr{G})$ with $\hat{\psi}_{\mathrm{CL}}$ as an initial guess.

This new call to the stochastic algorithm is explained by the fact that $\pi_{\mathrm{CL}}(\cdot \mid \mathbf{x}, \mathscr{G})$ and $\pi_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right)$ do not share the same maximum *a posteriori*. Indeed the magnitude and the curvature adjustments involve the prior on parameter ψ in their construction. Hence, once we plug the weight in $\pi_{\mathrm{CL}}^{\mathrm{cal}}$ the prior induces a bias between $\hat{\psi}_{\mathrm{CL}}$ and $\hat{\psi}_{\mathrm{cal}}$ aside from special cases where the maximum *a posteriori* $\hat{\psi}_{\mathrm{CL}}$ is a local extremum for the prior. As an example, concentrate on the magnitude adjustment even though the argument stays in essence the same for the curvature adjustment. The composite posterior distribution can be written as

$$\begin{aligned}\pi_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right) &\propto f_{\mathrm{CL}}\left(\mathbf{x} \,\middle|\, \psi, \mathscr{G}\right)^{w} \pi(\psi) \\ &\propto \pi_{\mathrm{CL}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right)^{w} \pi(\psi)^{1-w}.\end{aligned}$$

It follows that

$$\nabla \log \pi_{\mathrm{CL}}^{\mathrm{cal}}\left(\hat{\psi}_{\mathrm{CL}} \,\middle|\, \mathbf{x}, \mathscr{G}\right) = (1-w)\, \nabla \log \pi\left(\hat{\psi}_{\mathrm{CL}}\right),$$

which is generally non-zero.

Note there is a difference between adjusting the composite likelihood and the composite posterior difference. We chose to correct the composite posterior distribution instead of plugging in an adjusted composite likelihood since the latter possibility does not guarantee that the resulting composite posterior and posterior distributions would have the maximum *a posteriori*. Even if the techniques are the same, in addition

of the maximum composite likelihood estimator $\hat{\psi}_{\text{MCLE}}$ and the maximum likelihood estimator $\hat{\psi}_{\text{MLE}}$, one has to estimate $\hat{\psi}_{\text{cal}}$ and $\hat{\psi}_{\text{MAP}}$ which is the most cumbersome part of the procedure.

2.3 Examples

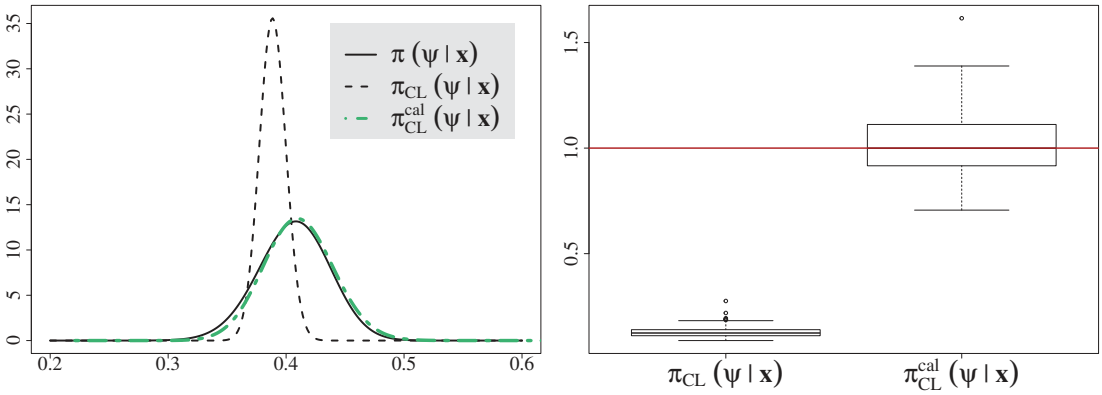
In this numerical part of the paper, we focus on models defined on a 16×16 lattice and we use exhaustively all 4×4 blocks. For the lattice of this dimension the recursions proposed by Friel and Rue (2007) can be used to compute exactly the normalizing constants $Z(\psi, \mathscr{G})$, $Z(\theta, \mathscr{G}, \mathbf{x}_{A(i)})$ and to draw exactly from the distribution $\pi(\cdot \mid \psi, \mathscr{G})$ or from the full-conditional distribution of blocks $A(i)$, namely $\pi(\cdot \mid \mathbf{x}_{-A(i)}, \psi, \mathscr{G})$. This exact computation of the posterior serves as a ground truth against which to compare with the posterior estimates of ψ using the various composite likelihood estimators. Computation was carried out on a desktop PC with six 3.47Ghz processors and with 8Gb of memory. Computing the normalizing constant of each block took 0.0004 second of CPU time. One iteration of the BFGS algorithm took 0.09 seconds to estimate the MAP of the composite likelihood and 1 second to estimate the MAP of true likelihood. The weight calibration for one dataset took approximately four minutes. Note that for more realistic situations involving larger lattices, one requires a sampler to draw from the full likelihood such as the Swendsen-Wang algorithm (Swendsen and Wang, 1987), however the computational cost of using this algorithm increases dramatically with the size of the lattice. One possible alternative is the slice sampler of Mira et al. (2001) that provides exact simulations of Ising models.

In all experiments, we simulated 100 realisations from the model and we placed uniform priors on ψ. For each realisation, we used the BFGS algorithm described in Section 2.2.3 with an adhoc stopping condition to get the estimators $\hat{\psi}_{\text{MAP}}$ and $\hat{\psi}_{\text{CL}}$. One iteration of the algorithm was based on a Monte Carlo estimator of either $\mathbf{E}_{\psi}\{\mathbf{S}(\mathbf{X})\}$ or $\mathbf{E}_{\psi}\left\{\mathbf{S}(\mathbf{X}_{A(i)} \mid \mathbf{x}_{-A(i)}, \psi, \mathscr{G})\right\}$ calculated from 100 exact draws whereas the Monte Carlo estimators of the covariance matrix $\mathbf{Var}_{\hat{\psi}_{\text{MAP}}}\{\mathbf{S}(\mathbf{X})\}$ and $\mathbf{Var}_{\hat{\psi}_{\text{CL}}}\left\{\mathbf{S}(\mathbf{X}_{A(i)} \mid \mathbf{x}_{-A(i)})\right\}$ used to compute the different weights were based on 50000 exact draws.

When the parameter space lies in $\mathbb{R}^2$, system of equations can be easily solved and the curvature adjustment was considered for all possible lower triangular matrices solutions. For the examples below we have four possible choices for W. Denote

$$W = \begin{pmatrix} w_{11} & 0 \\ w_{21} & w_{22} \end{pmatrix}$$

a solution of (2.9). Then possible solutions $\left\{W^{(1)}, W^{(2)}, W^{(3)}, W^{(4)}\right\}$ are set such that



(a) (b)

Figure 2.2: First experiment results. (a) Posterior parameter distribution (plain), non-calibrated composite posterior parameter distribution (dashed) and composite posterior distribution (green) of a first-order Ising model. (b) Boxplot displaying the ratio of the variance of the composite posterior parameter by the variance of the posterior parameter for 100 realisations of a first-order Ising model.

- $\det\left(W^{(1)}\right) < 0$ and $w_{21} > 0$,
- $\det\left(W^{(2)}\right) > 0$ and $w_{21} > 0$,
- $\det\left(W^{(3)}\right) < 0$ and $w_{21} < 0$,
- $\det\left(W^{(4)}\right) > 0$ and $w_{21} < 0$.

As a mean to assess the performance of the various adjustments, we propose to compare the posterior covariance matrices for ψ. The latter are computed using numerical integration methods which are detailed below. Denote $\mathbf{Var}_{\mathrm{CL}}\{\psi\}$ and $\mathbf{Var}\{\psi\}$ the variance of the composite posterior parameter and the variance of the posterior parameter respectively. In particular, we evaluate the relative mean square error

$$\mathrm{RMSE} = \mathbf{E}\left[\left\|1 - \mathbf{Var}_{\mathrm{CL}}(\psi)\mathbf{Var}^{-1}(\psi)\right\|_{\mathrm{F}}^2\right],$$

where $\|\cdot\|_{\mathrm{F}}$ it the Frobenius norm. When it is relevant, we also report the expected Kullback-Leibler divergence between the composite posterior and true posterior distributions

$$\mathrm{EKLD} = \mathbf{E}\left\{\mathrm{KL}\left(\pi^{\mathrm{cal}}_{\mathrm{CL}}, \pi(\cdot \mid \mathbf{X}, \mathscr{G})\right)\right\}.$$

First experiment We considered a first-order Ising model with a single interaction parameter $\psi \equiv \beta = 0.4$, which is close to the critical phase transition beyond which all realised lattices takes either value +1 or -1. This parameter setting is the most

challenging for the Ising model, since realised lattices exhibit strong spatial correlation around this parameter value. Using a fine grid of $\{\psi_k\}$ values, the right hand side of:

$$\pi\left(\psi_k \,\middle|\, \mathbf{x}, \mathscr{G}\right) \propto \pi\left(\mathbf{x} \,\middle|\, \psi_k, \mathscr{G}\right) \pi\left(\psi_k\right), \; k = 1, \ldots, N,$$

can be evaluated exactly. Summing up the right hand side – using the trapezoidal rule – yields an estimate of the evidence, $\pi(\mathbf{x})$, which is the normalizing constant for the expression above and which in turn can be used to give a very precise estimate of $\pi(\psi_k \mid \mathbf{x}, \mathscr{G})$. The posterior variance of ψ and the Kullback-Leibler divergence is estimated with trapezoidal rule on the same grid.

The plot so obtained for the posterior distribution and composite posterior distribution are given by Figure 2.2(a). On this example it should be clear that using an non-calibrated conditional composite likelihood leads to considerably underestimated posterior variances. But once we perform the mode adjustment and the magnitude adjustment, this provides a very good approximation of the true posterior. In Figure 2.2(b) we display the ratio of the variance of the composite posterior parameter by the variance of the posterior parameter based on $n_{\text{obs}} = 100$ realisations of a first-order Ising model. In view of these results there is no question that the magnitude adjustment (2.8) provides an efficient correction of the variance. Table 2.2 fills in these empirical results through evaluation of the relative mean square error (RMSE) and the expected Kullback-Leibler divergence (EKLD) between the composite posterior and true posterior distributions.

Table 2.2: Evaluation of the relative mean square error (RMSE) and the expected KL-divergence (EKLD) between the approximated posterior and true posterior distributions for 100 simulations of a first-order Ising model in the first experiment.

Composite posterior distribution	RMSE	EKLD
$\pi_{\text{CL}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right)$	0.870	0.337
$\overline{\pi}_{\text{CL}}^{\text{cal}}\left(\psi \,\middle|\, \mathbf{x}, \mathscr{G}\right)$	0.021	0.010

Second experiment We were interested in an anisotropic configuration of a first-order Ising model. We set $\psi = (\beta_1, \beta_2) = (0.3, 0.5)$ (see Table 1.1). The numerical integration is performed using an unstructured grid of triangles on the domain of the prior. Overall, the adjustments perform very well and as for the isotropic case, the mode and the magnitude adjustment allows us to build an accurate approximation of the posterior. Figure 2.3(a) and Figure 2.3(b) represent a comparison between the posterior distribution and the composite posterior distribution. In Figure 2.3(c) we display boxplots of the ratio $\frac{1}{\sqrt{2}}\|\mathbf{Var}_{\text{CL}}(\psi)\mathbf{Var}^{-1}(\psi)\|_{\text{F}}$, where $\|\cdot\|_{\text{F}}$ denotes the

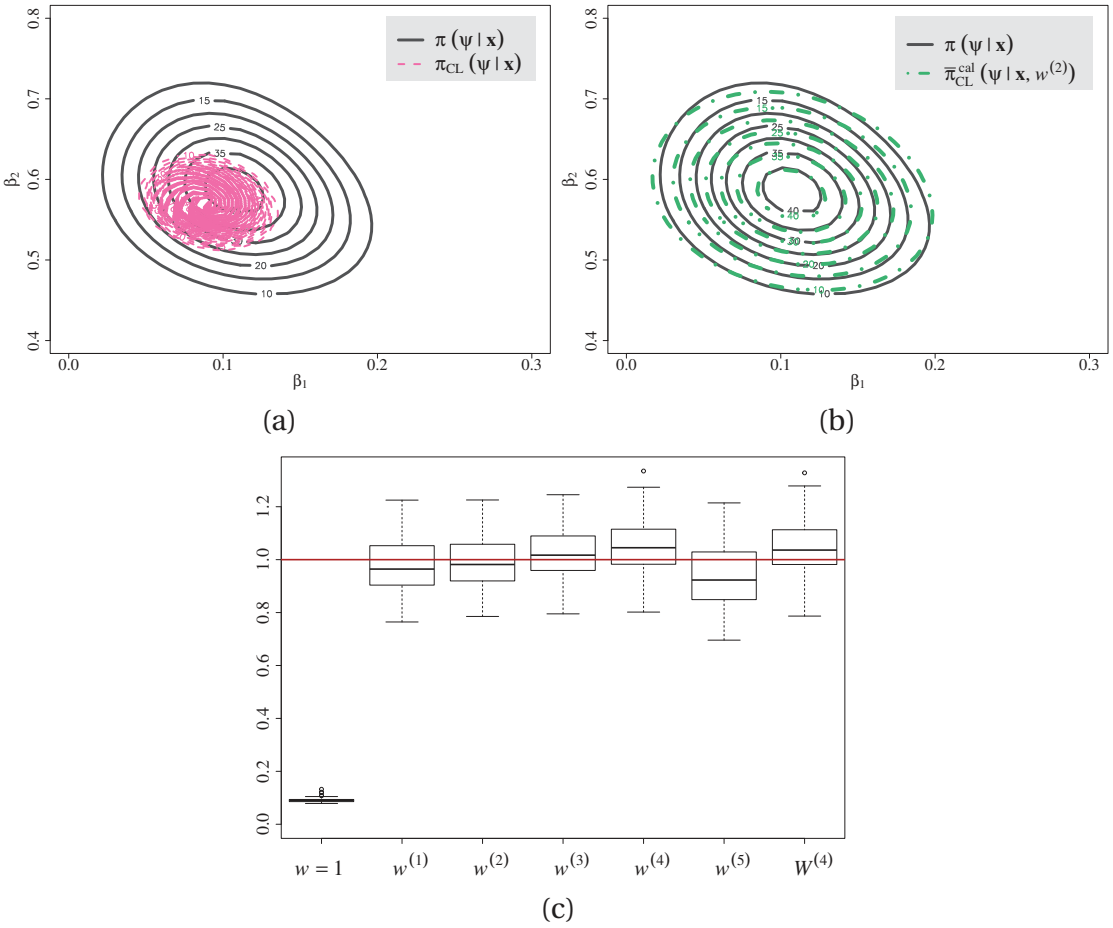


Figure 2.3: Second experiment results. (a) Posterior parameter distribution (grey) and non-calibrated composite posterior parameter distribution (pink) for a first-order anisotropic Ising model. (b) Posterior parameter distribution (grey) and composite posterior parameter distribution (green) with mode and magnitude adjustments ($w = w^{(2)}$). (c) Boxplots displaying $\frac{1}{\sqrt{2}}\|\mathbf{Var}_{\mathrm{CL}}(\theta)\mathbf{Var}^{-1}(\theta)\|_{\mathrm{F}}$ for 100 realisations of an anisotropic first-order Ising model.

Frobenius norm, for 100 realisations of an anisotropic first-order Ising model. The different weight options are almost equivalent in term of variance correction even if it seems that the Frobenius norm (weight $w^{(5)}$) leads to somewhat underestimate the posterior variance.

These observations can be further discussed with Table 2.3 that presents the relative mean square error and the average KL-divergence between the composite posterior distribution and the posterior distribution for 100 realisations of the model. As regards the magnitude adjustment, whilst the RMSE is little lower for weights $w^{(1)}$ and $w^{(2)}$, the performance of the composite posterior distributions are significantly the same in term of the Kullback-Leibler divergence. By contrast, we can observe that the curva-

Table 2.3: Evaluation of the relative mean square error (RMSE) the expected KL-divergence (EKLD) between the composite posterior distribution and true posterior distribution for 100 simulations of an anisotropic first-order Ising model.

Composite posterior distribution	RMSE	EKLD
$\pi_{\mathrm{CL}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right)$	1.28	1.68
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right)$ with $w = w^{(1)}$	0.269	0.044
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right)$ with $w = w^{(2)}$	0.265	0.042
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right)$ with $w = w^{(3)}$	0.272	0.042
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right)$ with $w = w^{(4)}$	0.285	0.043
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right)$ with $w = w^{(5)}$	0.283	0.047
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}, W^{(4)}\right)$	0.266	0.038

ture adjusment for the same RMSE, yields better result regarding the Kullback-Leibler divergence. The use of the curvature adjustment allows to reduce the importance of local dissimilarities and to capture more of the distribution tails.

Third experiment Here we focused on an autologistic model with a first-order dependence structure. The abundance parameter was set to $\alpha = 0.05$ and the interaction parameter to $\beta = 0.4$. The different implementations settings are exactly the same as for the second experiment. This example illustrates how the use of composite posterior distribution can induce a modification of the geometry of the distribution as shown in Figure 2.4(a).

Indeed in addition to the mode and variance misspecifications the conditional composite likelihood also changes the correlation between the variables. It should be evident that a magnitude adjustment would not be fruitful here since it would not

Table 2.4: Evaluation of the relative mean square error (RMSE) for 100 simulations of a first-order autologistic model.

Composite posterior distribution	RMSE
$\pi_{\mathrm{CL}}\left(\psi \mid \mathbf{x}, \mathscr{G}\right)$	1.19
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}, W^{(1)}\right)$	0.86
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}, W^{(2)}\right)$	0.89
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}, W^{(3)}\right)$	0.77
$\overline{\pi}_{\mathrm{CL}}^{\mathrm{cal}}\left(\psi \mid \mathbf{x}, \mathscr{G}, W^{(4)}\right)$	0.69

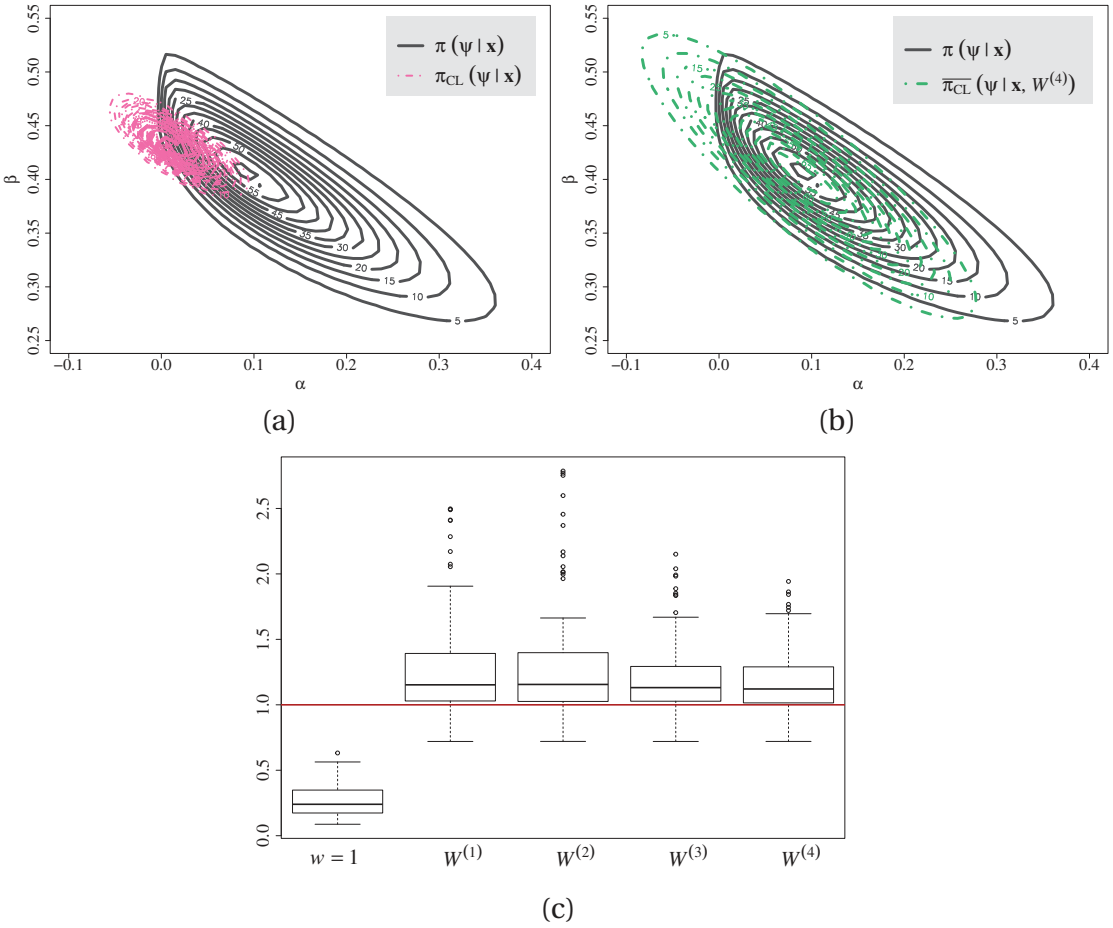


Figure 2.4: Third experiment results. (a) Posterior parameter distribution (grey) and non-calibrated composite posterior parameter distribution (pink) for a first-order autologistic model. (b) Posterior parameter distribution (grey) and composite posterior parameter distribution (green) with mode and curvature adjustments ($W = W^{(4)}$). (c) Boxplots displaying $\frac{1}{\sqrt{2}}\|\mathbf{Var}_{\mathrm{CL}}(\psi)\mathbf{Var}^{-1}(\psi)\|_{\mathrm{F}}$ for 100 realisations of a first-order autologistic model.

affect the correlation. Instead the curvature adjustment manages to do so and thus yields a good approximation of the posterior, see Figure 2.4(b). Overall adjusted composite posterior distribution perform better than the somewhat crude non-calibrated surrogate. We observe on Figure 2.4(c) a trend to overestimate the posterior variance of model parameter. This observation comes from we do not detect posterior tails (Figure 2.4(b)). In addition to the latter observations Table 2.4 shows the efficiency in terms of the RMSE and concludes that the best performances are obtained with $W = W^{(4)}$.

3 ABC model choice for hidden Gibbs random fields

Selecting between different dependency structures of hidden Markov random field can be very challenging, due to the intractable normalizing constant in the likelihood. In this chapter, we answer this question with approximate Bayesian computation (ABC, Tavaré et al., 1997, Pritchard et al., 1999, Marin et al., 2012, Baragatti and Pudlo, 2014) which provides a model choice method in the Bayesian paradigm. This comes after the work of Grelaud et al. (2009) who exhibited sufficient statistics on directly observed Gibbs random fields. As pointed in Section 1.8.2 this property is a peculiarity of models whose potential linearly depends on parameters and the sufficiency falls in the hidden case. This raises major difficulties that have been especially highlighted for model choice (Robert et al., 2011, Didelot et al., 2011). 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 work 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.

Recent articles have proposed automatic schemes to construct theses 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. But very few has been accomplished in the context of ABC model choice apart from the work of Prangle et al. (2014). The statistics $\mathbf{S}(\mathbf{y})$ 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.

We propose to complement the set with geometric summary statistics. The general approach to construct these intuitive statistics relies on a clustering analysis of the sites based on the observed colors and plausible latent graphs.

The Chapter is organized as follows: Section 3.1 presents ABC model choice as a k-nearest neighbor classifier, and defines a local error rate which is the first contribution of the Chapter. As a byproduct we provide an ABC algorithm based on the local error to select automatically the dimension of the summary statistics without distorting the model selection. The second contribution is the introduction of a general and intuitive approach to produce geometric summary statistics for hidden Potts model in Section 3.4. We end the Chapter with numerical results in that framework.

3.1 Local error rates and adaptive ABC model choice

When dealing with models whose likelihood cannot be computed analytically, Bayesian model choice becomes challenging since the evidence of each model writes as the integral of the likelihood over the prior distribution of the model parameter (see Section 1.8.1). ABC provides a method to escape from the intractability problem and relies on many simulated datasets from each model either to learn the model that fits the observed data $\mathbf{y}^{\text{obs}}$ or to approximate the posterior probabilities. We refer the reader to Section 1.8.2 and to reviews on ABC (Marin et al., 2012, Baragatti and Pudlo, 2014) to get a wider presentation.

3.1.1 Background on Approximate Bayesian computation for model choice

Recall the framework of Bayesian model selection introduced in Section 1.8.1. Assume we are given a set $\mathscr{M} = \{m : 1, \ldots, M\}$ 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 $\mathscr{M}$, $\pi(1), \ldots, \pi(M)$,

(ii) for each model m, a prior on its parameter space Θ_m, whose density is $\pi_m(\theta_m)$ and

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

The evidence of model m is then defined as

$$e(\mathbf{y} \mid m) = \int \pi_m(\mathbf{y} \mid \theta_m)\, \pi_m(\theta_m)\, \mathrm{d}\theta_m.$$

and the posterior probability of model m as

$$\pi\left(m \,\middle|\, \mathbf{y}\right) = \frac{\pi(m)e\left(\mathbf{y} \,\middle|\, m\right)}{\sum_{m'} \pi(m')e\left(\mathbf{y} \,\middle|\, m'\right)}. \tag{3.1}$$

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

$$\widehat{m}_{\text{MAP}}\left(\mathbf{y}^{\text{obs}}\right) = \arg\max_{m} \pi\left(m \,\middle|\, \mathbf{y}^{\text{obs}}\right). \tag{3.2}$$

The latter can be seen as a classification problem predicting the model number given the observation of $\mathbf{y}$. From this standpoint, $\widehat{m}_{\text{MAP}}$ is the Bayes classifier, well known to minimize the 0-1 loss (Devroye et al., 1996). One might argue that $\widehat{m}_{\text{MAP}}$ is an estimator defined as the mode of the posterior probabilities which form the density of the posterior with respect to the counting measure. But the counting measure, namely $\delta_1 + \cdots + \delta_M$, is a canonical reference measure, since it is invariant to any permutation of $\{1,\ldots,M\}$ whereas no such canonical reference measure (invariant to one-to-one transformation) exists on compact subset of the real line. Thus (3.2) does not suffer from the drawbacks of posterior mode estimators (Druilhet and Marin, 2007).

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 (3.2): it approximates the posterior probabilities by the frequency of each model number associated with simulated $\mathbf{y}$'s in a neighborhood of $\mathbf{y}^{\text{obs}}$. If required, we can eventually predict the best model with the most frequent model in the neighborhood, or, in other words, take the final decision by plugging in (3.2) the approximations of the posterior probabilities.

If directly applied, this first, naive algorithm faces the curse of dimensionality, as simulated datasets $\mathbf{y}$ can be complex objects and lie in a space of high dimension (*e.g.*, numerical images). Indeed, finding a simulated dataset in the vicinity of $\mathbf{y}^{\text{obs}}$ is almost impossible when the ambient dimension is high. The ABC algorithm performs therefore a (non linear) projection of the observed and simulated datasets onto some Euclidean space of reasonable dimension via a function $\mathbf{S}$, composed of summary statistics. Moreover, due to obvious reasons regarding computer memory, instead of keeping track of the whole simulated datasets, one commonly saves only the simulated vectors of summary statistics, which leads to a table composed of *iid* replicates $\{m,\theta_m,\mathbf{S}(\mathbf{y})\}$, often called the reference table in the ABC literature, see Algorithm 9.

Algorithm 9: Simulation of the ABC reference table

Output: A reference table of size n_{REF}

for $j \leftarrow 1$ **to** n_{REF} **do**
  draw m from the prior π;
  draw θ from the prior π_m;
  draw $\mathbf{y}$ from the likelihood $\pi_m(\cdot \mid \theta)$;
  compute $\mathbf{S}(\mathbf{y})$;
  save $\left\{m^{(j)}, \theta^{(j)}, \mathbf{S}\left(\mathbf{y}^{(j)}\right)\right\} \leftarrow \{m, \theta, \mathbf{S}(\mathbf{y})\}$;
end

return the table of $\left\{m^{(j)}, \theta^{(j)}, \mathbf{S}\left(\mathbf{y}^{(j)}\right)\right\}$, $j = 1, \ldots, n_{\text{REF}}$

From the standpoint of machine learning, the reference table serves as a training database composed of *iid* replicates drawn from the distribution of interest, namely the joint Bayesian model. The regression problem of estimating the posterior probabilities or the classification problem of predicting a model number are both solved with non-parametric methods. The neighborhood of $\mathbf{y}^{\text{obs}}$ is thus defined as simulations whose distances to the observation measured in terms of summary statistics, *i.e.*, $\rho\left\{\mathbf{S}(\mathbf{y}), \mathbf{S}\left(\mathbf{y}^{\text{obs}}\right)\right\}$, fall below a threshold ε commonly named the tolerance level. The calibration of ε is delicate, but had been partly neglected in the papers dealing with ABC that first focused on decreasing the total number of simulations via the recourse to Markov chain Monte Carlo (Marjoram et al., 2003) or sequential Monte Carlo methods (Beaumont et al., 2009, Del Moral et al., 2012) whose common target is the joint Bayesian distribution conditioned by $\rho\left\{\mathbf{S}(\mathbf{y}), \mathbf{S}\left(\mathbf{y}^{\text{obs}}\right)\right\} \leq \varepsilon$ for a given ε. By contrast, the simple setting we adopt here reveals the calibration question. In accepting the machine learning viewpoint, we can consider the ABC algorithm as a k-nearest neighbor (knn) method, see Biau et al. (2013); the calibration of ε is thus transformed into the calibration of k. The Algorithm we have to calibrate is given in Algorithm 10.

Before entering into the tuning of k, we highlight that the projection via the summary statistics generates a difference with the standard knn methods. Under mild conditions, knn are consistent non-parametric methods. Consequently, as the size of the reference table tends to infinity, the relative frequency of model m returned by Algorithm 10 converges to

$$\pi\left(m \,\middle|\, S(\mathbf{y}^{\text{obs}})\right).$$

Unfortunately, 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

Algorithm 10: Uncalibrated ABC model choice

Output: A sample of size k distributed according to the ABC approximation of the posterior

simulate the reference table $\mathscr{T}$ according to Algorithm 9;

sort the replicates of $\mathscr{T}$ according to $\rho\left\{\mathbf{S}\left(\mathbf{y}^{(j)}\right), \mathbf{S}\left(\mathbf{y}^{\text{obs}}\right)\right\}$;

keep the k first replicates;

return the relative frequencies of each model among the k first replicates and the most frequent model;

can greatly differ from the genuine $\pi(m \mid \mathbf{y}^{\text{obs}})$. Afterwards Marin et al. (2014) provide necessary and sufficient conditions on $\mathbf{S}(\cdot)$ for the consistency of the MAP based on $\pi\left(m \mid \mathbf{S}(\mathbf{y}^{\text{obs}})\right)$ when the information included in the dataset $\mathbf{y}^{\text{obs}}$ increases, *i.e.* when the dimension of $\mathbf{y}^{\text{obs}}$ tends to infinity. Consequently, the problem that ABC addresses with reliability is classification, and the mentioned theoretical results requires a shift from the approximation of posterior probabilities. Practically the frequencies returned by Algorithm 10 should solely be used to order the models with respect to their fits to $\mathbf{y}^{\text{obs}}$ and construct a knn classifier $\widehat{m}$ that predicts the model number.

It becomes therefore obvious that the calibration of k should be done by minimizing the misclassification error rate of the resulting classifier $\widehat{m}$. This indicator is the expected value of the 0-1 loss function, namely $\mathbf{1}\left\{\widehat{m}(\mathbf{y}) \neq m\right\}$, over a random $(m, \mathbf{y})$ distributed according to the marginal (integrated in θ_m) of the joint Bayesian distribution, whose density in $(m, \mathbf{y})$ writes

$$\pi(m) \int \pi_m\left(\mathbf{y} \mid \theta_m\right) \pi_m\left(\theta_m\right) \mathrm{d}\theta_m. \tag{3.3}$$

Ingenious solutions have been already proposed and are now well established to fulfil this minimization goal and bypass the overfitting problem, based on cross-validation on the learning database. But, for the sake of clarity, particularly in the following sections, we decided to take advantage of the fact that ABC aims at learning on simulated databases, and we will use a validation reference table, simulated also with Algorithm 9, but independently of the training reference table, to evaluate the misclassification rate with the averaged number of differences between the true model numbers $m^{(j)}$ and the predicted values $\widehat{m}(\mathbf{y}^{(j)})$ by knn (*i.e.*, by ABC) on the validation reference table.

3.1.2 Local error rates

The misclassification rate τ of the knn classifier $\hat{m}$ at the core of Algorithm 10 provides consistent evidence of its global accuracy. It supplies indeed a well-known support to calibrate k in Algorithm 10. The purpose of ABC model choice methods though is the analyse of an observed dataset $\mathbf{y}^{\text{obs}}$ and this first indicator is irrelevant to assess the accuracy of the classifier at this precise point of the data space, since it is by nature a prior gauge. We propose here to disintegrate this indicator, and to rely on conditional expected value of the misclassification loss $\mathbf{1}\left\{\hat{m}(\mathbf{y}) \neq m\right\}$ knowing $\mathbf{y}$ as an evaluation of the efficiency of the classifier at $\mathbf{y}$. We recall the following proposition whose proof is easy, but might help clarifying matters when applied to the joint distribution (3.3).

Proposition 3.2. *Consider a classifier $\hat{m}$ that aims at predicting m given $\mathbf{y}$ on data drawn from the joint distribution $f(m,\mathbf{y})$. Let τ be the misclassification rate of $\hat{m}$, defined by $\mathbf{P}(\hat{m}(\mathbf{Y}) \neq \mathscr{M})$, where $(\mathscr{M},\mathbf{Y})$ is a random pair with distribution f under the probability measure $\mathbf{P}$. Then,* (i) *the expectation of the loss function is*

$$\tau = \sum_m \int_{\mathscr{Y}} \mathbf{1}\left\{\hat{m}(\mathbf{y}) \neq m\right\} f(m,\mathbf{y})\,\mathrm{d}\mathbf{y}.$$

Additionally, (ii)*, the conditional expectation knowing $\mathbf{y}$, namely*

$$\tau(\mathbf{y}) = \mathbf{P}\left(\hat{m}(\mathbf{Y}) \neq \mathscr{M} \,\middle|\, \mathbf{Y} = \mathbf{y}\right),$$

is

$$\tau(\mathbf{y}) = \sum_m \mathbf{1}\left\{\hat{m}(\mathbf{y}) \neq m\right\} f(m \mid \mathbf{y}) \tag{3.4}$$

and $\tau = \int_{\mathscr{Y}} f(\mathbf{y})\tau(\mathbf{y})\,\mathrm{d}\mathbf{y}$, where $f(\mathbf{y})$ denotes the marginal distribution of f (integrated over m) and $f(m \mid \mathbf{y}) = f(m,\mathbf{y})/f(\mathbf{y})$ the conditional probability of m given $\mathbf{y}$. Furthermore, we have

$$\tau(\mathbf{y}) = 1 - f(\hat{m}(\mathbf{y}) \mid \mathbf{y}). \tag{3.5}$$

The last result (3.5) suggests that a conditional expected value of the misclassification loss is a valuable indicator of the error at $\mathbf{y}$ since it is admitted that the posterior probability of the predicted model reveals the accuracy of the decision at $\mathbf{y}$. Nevertheless, the whole simulated datasets are not saved into the ABC reference table but solely some numerical summaries $\mathbf{S}(\mathbf{y})$ per simulated dataset $\mathbf{y}$, as explained above. Thus the disintegration process of τ is practically limited to the conditional expectation of the loss knowing some non one-to-one function of $\mathbf{y}$. Its definition becomes therefore

much more subtle than the basic (3.4). Actually, the ABC classifier can be trained on a subset $\mathbf{S}_1(\mathbf{y})$ of the summaries $\mathbf{S}(\mathbf{y})$ saved in the training reference table, or on some deterministic function (we still write $\mathbf{S}_1(\mathbf{y})$) of $\mathbf{S}(\mathbf{y})$ that reduces the dimension, such as the projection on the LDA axes proposed by Estoup et al. (2012). To highlight this fact, the ABC classifier is denoted by $\widehat{m}(\mathbf{S}_1(\mathbf{y}))$ in what follows. It is worth noting here that the above setting encompasses any dimension reduction technique presented in the review of Blum et al. (2013), though the review in oriented on parameter inference. Furthermore we might want to disintegrate the misclassification rate with respect to another projection $\mathbf{S}_2(\mathbf{y})$ of the simulated data that can or cannot be related to the summaries $\mathbf{S}_1(\mathbf{y})$ used to train the ABC classifier, albeit $\mathbf{S}_2(\mathbf{y})$ is also limited to be a deterministic function of $\mathbf{S}(\mathbf{y})$. This yields the following.

Definition 5. *The local error rate of the* $\widehat{m}(\mathbf{S}_1(\mathbf{y}))$ *classifier with respect to* $\mathbf{S}_2(\mathbf{y})$ *is*

$$\tau_{\mathbf{S}_1}(\mathbf{S}_2(\mathbf{y})) := \mathbf{P}\left(\widehat{m}(\mathbf{S}_1(\mathbf{Y})) \neq \mathscr{M} \,\middle|\, \mathbf{S}_2(\mathbf{Y}) = \mathbf{S}_2(\mathbf{y})\right),$$

where $(\mathscr{M}, \mathbf{Y})$ *is a random variable with distribution given in* (3.3).

The purpose of the local misclassification rate in the present Chapter is twofold and requires to play with the distinction between $\mathbf{S}_1$ and $\mathbf{S}_2$, as the last part will show on numerical examples. The first goal is the construction of a prospective tool that aims at checking whether a new statistic $\mathbf{S}'(\mathbf{y})$ carries additional information regarding the model choice, beyond a first set of statistics $\mathbf{S}_1(\mathbf{y})$. In the latter case, it can be useful to localize the misclassification error of $\widehat{m}(\mathbf{S}_1(\mathbf{y}))$ with respect to the concatenated vector $\mathbf{S}_2(\mathbf{y}) = \left(\mathbf{S}_1(\mathbf{y}), \mathbf{S}'(\mathbf{y})\right)$. Indeed, this local error rate can reveal concentrated areas of the data space, characterized in terms of $\mathbf{S}_2(\mathbf{y})$, in which the local error rate rises above $(M-1)/M$, the averaged (local) amount of errors of the random classifier among M models, so as to approach 1. The interpretation of the phenomenon is as follows: errors committed by $\widehat{m}(\mathbf{S}_1(\mathbf{y}))$, that are mostly spread on the $\mathbf{S}_1(\mathbf{y})$-space, might gather in particular areas of subspaces of the support of $\mathbf{S}_2(\mathbf{y}) = \left(\mathbf{S}_1(\mathbf{y}), \mathbf{S}'(\mathbf{y})\right)$. This peculiarity is due to the dimension reduction of the summary statistics in ABC before the training of the classifier and represents a concrete proof of the difficulty of ABC model choice already raised by Didelot et al. (2011) and Robert et al. (2011).

The second goal of the local error rate given in Definition 5 is the evaluation of the confidence we may concede in the model predicted at $\mathbf{y}^{\text{obs}}$ by $\widehat{m}(\mathbf{S}_1(\mathbf{y}))$, in which case we set $\mathbf{S}_2(\mathbf{y}) = \mathbf{S}_1(\mathbf{y})$. And, when both sets of summaries agree, the results of

Proposition 3.2 extend to

$$\begin{aligned}\tau_{\mathbf{S}_1}(\mathbf{S}_1(\mathbf{y})) &= \sum_m \pi\left(m \,\middle|\, \mathbf{S}_1(\mathbf{y})\right) \mathbf{1}\left\{\widehat{m}(\mathbf{S}_1(\mathbf{y})) = m\right\} \\ &= 1 - \pi\left(\widehat{m}(\mathbf{S}_1(\mathbf{y})) \,\middle|\, \S_1(\mathbf{y})\right). \end{aligned} \tag{3.6}$$

Besides the local error rate we propose in Definition 5 is an upper bound of the Bayes classifier if we admit the loss of information committed by replacing $\mathbf{y}$ with the summaries.

Proposition 3.3. *Consider any classifier $\widehat{m}(\mathbf{S}_1(\mathbf{y}))$. The local error rate of this classifier satisfies*

$$\begin{aligned}\tau_{\mathbf{S}_1}(\mathbf{s}_2) &= \mathbf{P}\left(\widehat{m}(\mathbf{S}_1(\mathbf{Y})) \neq \mathscr{M} \mid \mathbf{S}_2(\mathbf{Y}) = \mathbf{s}_2\right) \\ &\geq \mathbf{P}\left(\widehat{m}_{MAP}(\mathbf{Y}) \neq \mathscr{M} \mid \mathbf{S}_2(\mathbf{Y}) = \mathbf{s}_2\right), \end{aligned} \tag{3.7}$$

where $\widehat{m}_{MAP}$ is the Bayes classifier defined in (3.2) *and $\mathbf{s}_2$ any value in the support of $\mathbf{S}_2(\mathbf{Y})$. Consequently,*

$$\mathbf{P}\left(\widehat{m}(\mathbf{S}_1(\mathbf{Y})) \neq \mathscr{M}\right) \geq \mathbf{P}\left(\widehat{m}_{MAP}(\mathbf{Y}) \neq \mathscr{M}\right). \tag{3.8}$$

Proof. Proposition 3.2, in particular (3.5), implies that $\widehat{m}_{\mathrm{MAP}}(\mathbf{y})$ is the ideal classifier that minimizes the conditional 0-1 loss knowing $\mathbf{y}$. Hence, we have

$$\mathbf{P}\left(\widehat{m}(\mathbf{S}_1(\mathbf{Y})) \neq \mathscr{M} \,\middle|\, \mathbf{Y} = \mathbf{y}\right) \geq \mathbf{P}\left(\widehat{m}_{\mathrm{MAP}}(\mathbf{Y}) \neq \mathscr{M} \,\middle|\, \mathbf{Y} = \mathbf{y}\right).$$

Integrating the above with respect to the distribution of $\mathbf{Y}$ knowing $\mathbf{S}_2(\mathbf{Y})$ leads to (3.7), and a last integral to (3.8). □

Proposition 3.3 shows that the introduction of new summary statistics cannot distort the model selection insofar as the risk of the resulting classifier cannot decrease below the risk of the Bayes classifier $\widehat{m}_{\mathrm{MAP}}$. We give here a last flavour of the results of Marin et al. (2014) and mention that, if $\mathbf{S}_1(\mathbf{y}) = \mathbf{S}_2(\mathbf{y}) = \mathbf{S}(\mathbf{y})$ and if the classifiers are perfect (*i.e.*, trained on infinite reference tables), we can rephrase part of their results as providing mild conditions on $\mathbf{S}$ under which the local error $\tau_{\mathbf{S}}(\mathbf{S}(\mathbf{y}))$ tends to 0 when the size of the dataset $\mathbf{y}$ tends to infinity.

3.3.1 Estimation algorithm of the local error rates

The numerical estimation of the local error rate $\tau_{\mathbf{S}_1}(\mathbf{S}_2(\mathbf{y}))$, as a surface depending on $\mathbf{S}_2(\mathbf{y})$, is therefore paramount to assess the difficulty of the classification problem at any $\mathbf{s}_2 = \mathbf{S}_\mathbf{S}(\mathbf{y})$, and the local accuracy of the classifier. Naturally, when $\mathbf{S}_1(\mathbf{y}) = \mathbf{S}_2(\mathbf{y})$ for all $\mathbf{y}$, the local error can be evaluated at $\mathbf{S}_2(\mathbf{y}^{\text{obs}})$ by plugging in (3.6) the ABC estimates of the posterior probabilities (the relative frequencies of each model among the particles returned by Algorithm 10) as substitute for $\pi(m \mid \mathbf{S}(\mathbf{y}^{\text{obs}}))$. This estimation procedure is restricted to the above mentioned case where the set of statistics used to localize the error rate agrees with the set of statistics used to train the classifier. Moreover, the approximation of the posterior probabilities returned by Algorithm 10, *i.e.*, a knn method, might not be trustworthy: the calibration of k performed by minimizing the prior error rate τ does not provide any certainty on the estimated posterior probabilities beyond a ranking of these probabilities that yields the best classifier in terms of misclassification. In other words, the knn method calibrated to answer the classification problem of discriminating among models does not produce a reliable answer to the regression problem of estimating posterior probabilities. Certainly, the value of k must be increased to face this second kind of issue, at the price of a larger bias that might even swap the model ranking (otherwise, the empirical prior error rate would not depend on k, see the numerical result section).

For all these reasons, we propose here an alternative estimate of the local error. The core idea of our proposal is the recourse to a non-parametric method to estimate conditional expected values based on the calls to the classifier $\widehat{m}$ on a validation reference table, already simulated to estimate the global error rate τ. Nadaraya-Watson kernel estimators of the conditional expected values

$$\tau_{\mathbf{S}_1}(\mathbf{S}_2(\mathbf{y})) = \mathbf{E}\left(\mathbf{1}\left\{\widehat{m}(\mathbf{S}_1(\mathbf{Y})) \neq \mathscr{M}\right\} \,\middle|\, \mathbf{S}_2(\mathbf{Y}) = \mathbf{S}_2(\mathbf{y})\right) \tag{3.9}$$

rely explicitly on the regularity of this indicator, as a function of $\mathbf{s}_2 = \mathbf{S}_2(\mathbf{y})$, which contrasts with the ABC plug-in estimate described above. We thus hope improvements in the accuracy of error estimate and a more reliable approximation of the whole function $\tau_{\mathbf{S}_1}(\mathbf{S}_2(\mathbf{y}))$. Additionally, we are not limited anymore to the special case where $\mathbf{S}_1(\mathbf{y}) = \mathbf{S}_2(\mathbf{y})$ for all $\mathbf{y}$. It is worth stressing here that the bandwidth of the kernels must be calibrated by minimizing the L^2-loss, since the target is a conditional expected value.

Practically, this leads to Algorithm 11 which requires a *validation* or *test reference table* independent of the training database that constitutes the ABC reference table. We can bypass the requirement by resorting to cross validation methods, as for the computation of the global prior misclassification rate τ. But the ensued algorithm

Algorithm 11: Estimation of $\tau_{\mathbf{S}_1}(\mathbf{S}_2(\mathbf{y}))$ given an classifier $\widehat{m}(\mathbf{S}_1(\mathbf{y}))$ on a validation or test reference table

Input: A validation or test reference table and a classifier $\widehat{m}(\mathbf{S}_1(\mathbf{y}))$ fitted with a first reference table

Output: Estimations of (3.9) at each point of the second reference table

for each $\left(m^{(j)}, \mathbf{y}^{(j)}\right)$ in the test table **do**

  compute $\delta^{(j)} = \left\{\widehat{m}(\mathbf{S}_1(\mathbf{y}^{(j)})) \neq m^{(j)}\right\}$;

end

calibrate the bandwidth $\mathbf{h}$ of the Nadaraya-Watson estimator predicting $\delta^{(j)}$ knowing $\mathbf{S}_2(\mathbf{y}^{(j)})$ via cross-validation on the test table;

for each $\left(m^{(j)}, \mathbf{y}^{(j)}\right)$ in the test table **do**

  evaluate the Nadaraya-Watson estimator with bandwidth $\mathbf{h}$ at $\mathbf{S}_2\left(\mathbf{y}^{(j)}\right)$;

end

is complex and it induces more calls to the classifier (consider, *e.g.*, a ten-fold cross validation algorithm computed on more than one random grouping of the reference table) than the basic Algorithm 11, whereas the training database can always be supplemented by a validation database since ABC, by its very nature, is a learning problem on simulated databases. Moreover, to display the whole surface $\tau_{\mathbf{S}_1}(\mathbf{S}_2(\mathbf{y}))$, we can interpolate values of the local error between points $\mathbf{S}_2(\mathbf{y})$ of the second reference table with the help of a Kriging algorithm. We performed numerical experiments (not detailed here) concluding that the resort to a Kriging algorithm provides results comparable to the evaluation of Nadaraya-Watson estimator at any point of the support of $\mathbf{S}_2(\mathbf{y})$, and can reduce computation times.

3.3.2 Adaptive ABC

The local error rate can also represent a valuable way to adjust the summary statistics to the data point $\mathbf{y}$ and to build an adaptive ABC algorithm achieving a local trade off that increases the dimension of the summary statistics at $\mathbf{y}$ only when the additional coordinates add information regarding the classification problem. Assume that we have at our disposal a collection of ABC classifiers, $\widehat{m}_\lambda(\mathbf{y}) := \widehat{m}_\lambda(\mathbf{S}_\lambda(\mathbf{y}))$, $\lambda = 1, \ldots, \Lambda$, trained on various projections of $\mathbf{y}$, namely the $\mathbf{S}_\lambda(\mathbf{y})$'s, and that all these vectors, sorted with respect to their dimension, depend only on the summary statistics registered in the reference tables. Sometimes low dimensional statistics may suffice for the classification (of models) at $\mathbf{y}$, whereas other times we may need to examine statistics

of larger dimension. The local adaptation of the classifier is accomplished through the disintegration of the misclassification rates of the initial classifiers with respect to a common statistic $\mathbf{S}_0(\mathbf{y})$. Denoting $\tau_\lambda(\mathbf{S}_0(\mathbf{y}))$ the local error rate of $\widehat{m}_\lambda(\mathbf{y})$ knowing $\mathbf{S}_0(\mathbf{y})$, this reasoning yields the adaptive classifier defined by

$$\widetilde{m}(\mathbf{S}(\mathbf{y})) := \widehat{m}_{\widehat{\lambda}(\mathbf{y})}(\mathbf{y}), \text{ where } \widehat{\lambda}(\mathbf{y}) := \underset{\lambda=1,\ldots,\Lambda}{\arg\min}\, \tau_\lambda(S_0(\mathbf{y})). \tag{3.10}$$

This last classifier attempts to avoid bearing the cost of the potential curse of dimensionality from which all knn classifiers suffer and can help reduce the error of the initial classifiers, although the error of the ideal classifier (3.2) remains an absolute lower bound, see Proposition 3.3. From a different perspective, (3.10) represents a way to tune the similarity $\rho\left\{\mathbf{S}(\mathbf{y}), \S(\mathbf{y}^{\text{obs}})\right\}$ of Algorithm 10 that locally includes or excludes components of $\mathbf{S}(\mathbf{y})$ to assess the proximity between $\mathbf{S}(\mathbf{y})$ and $\mathbf{S}(\mathbf{y}^{\text{obs}})$. Practically, we rely on the following algorithm to produce the adaptive classifier, that requires a validation reference table independent of the reference table used to fit the initial classifiers.

Algorithm 12: Adaptive ABC model choice

Input: A collection of classifiers $\widehat{m}_\lambda(\mathbf{y})$, $\lambda = 1,\ldots,\Lambda$ and a validation reference table
Output: An adaptive classifier $\widetilde{m}(\mathbf{y})$

for each $\lambda \in \{1,\ldots,\Lambda\}$ **do**
  estimate the local error of $\widehat{m}_\lambda(\mathbf{y})$ knowing $S_0(y)$ with the help of Algorithm 11;
end

return the adaptive classifier $\widetilde{m}$ as a function computing (3.10);

The local error surface estimated within the loop of Algorithm 12 must contrast the errors of the collection of classifiers. Our advice is thus to build a projection $\mathbf{S}_0(\mathbf{y})$ of the summaries $\mathbf{S}(\mathbf{y})$ registered in the reference tables as follow. Add to the validation reference table a qualitative trait which groups the replicates of the table according to their differences between the predicted numbers by the initial classifiers and the model numbers $m^{(j)}$ registered in the database. For instance, when the collection is composed of $\Lambda = 2$ classifiers, the qualitative trait takes three values: value 0 when both classifiers $\widehat{m}_\lambda(\mathbf{y}^{(j)})$ agree (whatever the value of $\widehat{m}^{(j)}$), value 1 when the first classifier only returns the correct number, *i.e.*, $\widehat{m}_1(\mathbf{y}^{(j)}) = m^{(j)} \neq \widehat{m}_2(\mathbf{y}^{(j)})$, and value 2 when the second classifier only returns the correct number, *i.e.*, $\widehat{m}_1(\mathbf{y}^{(j)}) \neq m^{(j)} = \widehat{m}_2(\mathbf{y}^{(j)})$. The axes of the linear discriminant analysis (LDA) predicting the qualitative trait knowing $\mathbf{S}(\mathbf{y})$ provide a projection $\mathbf{S}_0(\mathbf{y})$ which contrasts the errors of the initial collection of classifiers.

Finally it is important to note that the local error rates are evaluated in Algorithm 12 with the help of a validation reference table. Therefore, a reliable estimation of the accuracy of the adaptive classifier cannot be based on the same validation database because of the optimism bias of the training error. Evaluating the accuracy requires the simulation of a *test reference table* independently of the two first databases used to train and adapt the predictor, as is usually performed in the machine learning community.

3.4 Hidden random fields

Our primary intent with the ABC methodology exposed in Section 3.1 was the study of new summary statistics to discriminate between hidden random fields models. The following materials numerically illustrate how ABC can choose the dependency structure of latent Potts models among two possible neighborhood systems, both described with undirected graphs, whilst highlighting the generality of the approach.

3.4.1 Hidden Potts model

This numerical part of the Chapter focuses on hidden Potts models, that are representative of the general level of difficulty while at the same time being widely used in practice (see for example Hurn et al., 2003, Alfò et al., 2008, François et al., 2006, Moores et al., 2014). We recall that the latent random field $\mathbf{x}$ is a family of random variables x_i indexed by a finite set $\mathscr{S}$ and taking values in a finite state space $\mathscr{X} = \{0, \ldots, K-1\}$. When modelling a digital image, the sites are lying on a regular 2D-grid of pixels, and their dependency is given by an undirected graph $\mathscr{G}$ which defines an adjacency relationship on the set of sites $\mathscr{S}$: by definition, both sites i and j are adjacent if and only if the graph $\mathscr{G}$ includes an edge that links directly i and j (see Chapter 1). A Potts model sets a probability distribution on $\mathbf{x}$, parametrized by a scalar β that adjusts the level of dependency between adjacent sites and defined by

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

We refer the reader to Section 1.1.2 for further details on Potss model.

In hidden Markov random fields, the latent process is observed indirectly through another field; this permits the modelling of a noise that may be encountered in many concrete situations. Precisely, given the realization $\mathbf{x}$ of the latent field, the observation $\mathbf{y}$ is a family of random variables indexed by the set of sites, and taking values in a set

$\mathscr{Y}$, *i.e.*, $\mathbf{y} = \{y_i : i \in \mathscr{S}\}$, and are commonly assumed as independent draws that form a noisy version of the hidden fields. Consequently, we set the conditional distribution of $\mathbf{y}$ knowing $\mathbf{x}$ as the product $\pi(\mathbf{y} \mid \mathbf{x}, \phi) = \prod_{i \in \mathscr{S}} \pi(y_i \mid x_i, \phi)$, where $\pi(y_i \mid x_i, \phi)$ is the marginal noise distribution parametrized by some scalar ϕ. Hence the likelihood of the hidden Potts model with parameter β on the graph $\mathscr{G}$ and noise distribution $\pi(\cdot \mid \mathbf{x}, \phi)$, denoted HPM$(\mathscr{G}, \phi, \beta)$, is given by

$$\pi\left(\mathbf{y} \,\middle|\, \phi, \beta, \mathscr{G}\right) = \sum_{\mathbf{x} \in \mathscr{X}} \pi\left(\mathbf{x} \,\middle|\, \beta, \mathscr{G}\right) \pi(\mathbf{y} \mid \mathbf{x}, \phi)$$

and faces a double intractable issue as neither the likelihood of the latent field, nor the above sum can be computed directly: the cardinality of the range of the sum is of combinatorial complexity. The following numerical experiments are based on two classes of noises, producing either observations in $\{0, 1, \ldots, K-1\}$, the set of latent colors, or continuous observations that take values in $\mathbb{R}$.

The common point of our examples is to select the hidden Gibbs model that better fits a given $\mathbf{y}^{\text{obs}}$ composed of $n = 100 \times 100$ pixels within different neighborhood systems represented as undirected graphs $\mathscr{G}$. We considered the two widely used adjacency structures in our simulations, namely the graph $\mathscr{G}_4$ (respectively $\mathscr{G}_8$) in which the neighborhood of a site is composed of the four (respectively eight) closest sites on the two-dimensional lattice, except on the boundaries of the lattice, see Fig. 1.1. The prior probabilities of both models were set to $1/2$ in all experiments. The Bayesian analysis of the model choice question adds another integral beyond the two above mentioned sums that cannot be calculated explicitly or numerically either and the problem we illustrate are said triple intractable. Up to our knowledge the choice of the latent neighborhood structure has never been seriously tackled in the Bayesian literature. We mentioned here the mean field approximation of Forbes and Peyrard (2003) whose software can estimate parameters of such models, and compare models fitness via a BIC criterion. This is discussed further in Chapter 4. The detailed settings of our three experiments are as follows.

First experiment. We considered Potts models with $K = 2$ colors and a noise process that switches each pixel independently with probability

$$\frac{\exp(-\phi)}{\exp(\phi) + \exp(-\phi)},$$

following the proposal of Everitt (2012). The prior on ϕ was uniform over $(0.42; 2.3)$, where the bounds of the interval were determined to switch a pixel with a probability less than 30%. Regarding the dependency parameter β, we set prior distributions

below the phase transition which occurs at different levels depending on the neighborhood structure. Precisely we used a uniform distribution over $(0;1)$ when the adjacency is given by $\mathscr{G}_4$ and a uniform distribution over $(0;0.35)$ with $\mathscr{G}_8$.

Second experiment. We increased the number of colors in the Potts models and set $K = 16$. Likewise, we set a noise that changes the color of each pixel with a given probability parametrized by ϕ, and conditionally on a change at site i, we rely on the least favourable distribution, which is a uniform draw within all colors except the latent one. To extend the parametrization of Everitt (2012), the marginal distribution of the noise is defined by

$$\pi\left(y_i \,\middle|\, x_i, \phi\right) = \frac{\exp\left\{\phi\left(2\mathbf{1}\left\{x_i = x_j - 1\right\}\right)\right\}}{\exp(\phi) + (K-1)\exp(-\phi)}$$

and a uniform prior on ϕ over the interval $(1.78;4.8)$ ensures that the probability of changing a pixel with the noise process is at most 30%. The uniform prior on the Potts parameter β was also tuned to stay below the phase transition. Hence β ranges the interval $(0;2.4)$ with a $\mathscr{G}_4$ structure and the interval $(0;1)$ with a $\mathscr{G}_8$ structure.

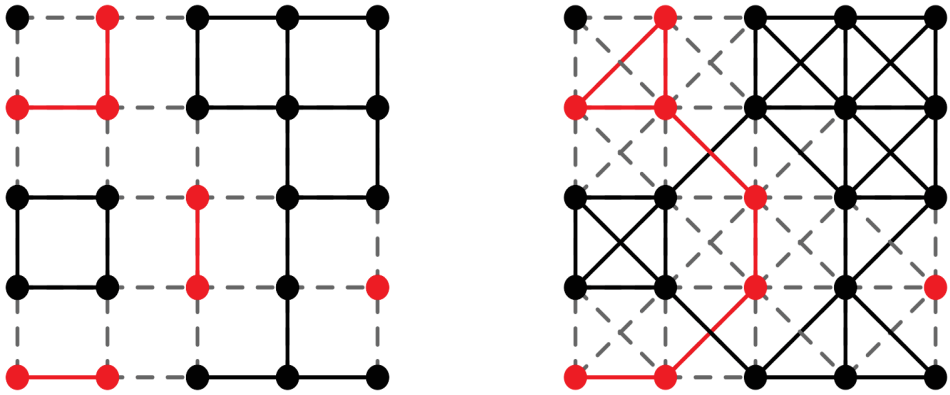
Third experiment. We introduced a homoscedastic Gaussian noise whose marginal distribution is characterized by

$$y_i \mid x_i = c \sim \mathcal{N}(c, \sigma^2) \quad c \in \{0;1\}$$

over bicolor Potts models. And both prior distributions on parameter β are similar to the ones on the latent fields of the first experiment. The standard deviation $\sigma = 0.39$ was set so that the probability of a wrong prediction of the latent color with a marginal MAP rule on the Gaussian model is about 10%.

3.4.2 Geometric summary statistics

Performing a Bayesian model choice via ABC algorithms requires summary statistics that capture the relevant information from the observation $\mathbf{y}^{\text{obs}}$ to discriminate among the competing models. When the observation is noise-free, Grelaud et al. (2009) noted that the joint distribution resulting from the Bayesian modelling falls into the exponential family, and they obtained consecutively a small set of summary statistics, depending on the collection of considered models, that were sufficient. In front of noise, the situation differs substantially as the joint distribution lies now outside the exponential family due to the bound to the data, and the above mentioned statistics



$\Gamma\left(\mathscr{G}_4,\mathbf{y}\right)$ $\Gamma\left(\mathscr{G}_8,\mathbf{y}\right)$

Figure 3.1: The induced graph $\Gamma\left(\mathscr{G}_4,\mathbf{y}\right)$ and $\Gamma\left(\mathscr{G}_8,\mathbf{y}\right)$ on a given bicolor image $\mathbf{y}$ of size 5×5. The six summary statistics on $\mathbf{y}$ are thus $R\left(\mathscr{G}_4,\mathbf{y}\right)=22$, $T\left(\mathscr{G}_4,\mathbf{y}\right)=7$, $U\left(\mathscr{G}_4,\mathbf{y}\right)=12$, $R\left(\mathscr{G}_8,\mathbf{y}\right)=39$, $T\left(\mathscr{G}_8,\mathbf{y}\right)=4$ and $U\left(\mathscr{G}_8,\mathbf{y}\right)=16$

are not sufficient anymore, whence the urge to bring forward other concrete and workable statistics. The general approach we developed reveals geometric features of a discrete field $\mathbf{y}$ via the recourse to coloured graphs attached to $\mathbf{y}$ and their connected components. Consider an undirected graph $\mathscr{G}$ whose set of vertices coincides with $\mathscr{S}$, the set of sites of $\mathbf{y}$.

Definition 6. *The graph induced by $\mathscr{G}$ on the field $\mathbf{y}$, denoted $\Gamma\left(\mathscr{G},\mathbf{y}\right)$, is the undirected graph whose set of edges gathers the edges of $\mathscr{G}$ between sites of $\mathbf{y}$ that share the same color, i.e.,*

$$i \overset{\Gamma(\mathscr{G},\mathbf{y})}{\sim} j \quad\Longleftrightarrow\quad i \overset{\mathscr{G}}{\sim} j \text{ and } y_i = y_j.$$

We believe that the connected components of such induced graphs capture major parts of the geometry of $\mathbf{y}$. Recall that a connected component of an undirected graph Γ is a subgraph of Γ in which any two vertices are connected to each other by a path, and which is connected to no other vertices of Γ. And the connected components form a partition of the vertices. Since ABC relies on the computation of the summary statistics on many simulated datasets, it is also worth noting that the connected components can be computed efficiently with the help of famous graph algorithms in linear time based on a breadth-first search or depth-first search over the graph. The empirical distribution of the sizes of the connected components represents an important source of geometric informations, but cannot be used as a statistic in ABC because of the curse of dimensionality. The definition of a low dimensional summary statistic derived from these connect components should be guided by the intuition on the model choice we face.

Our numerical experiments discriminate between a $\mathscr{G}_4$- and a $\mathscr{G}_8$-neighborhood structure and we considered two induced graphs on each simulated $\mathbf{y}$, namely $\Gamma\left(\mathscr{G}_4,\mathbf{y}\right)$ and $\Gamma\left(\mathscr{G}_8,\mathbf{y}\right)$. Remark that the two-dimensional statistics proposed by Grelaud et al. (2009), which are sufficient in the noise-free context, are the total numbers of edges in both induced graphs. After very few trials without success, we fixed ourselves on four additional summary statistics, namely the size of the largest component of each induced graph, as well as the total number of connect components in each graph. See Fig. 3.1 for an example on a bicolor picture $\mathbf{y}$. To fix the notations, for any induced graph $\Gamma\left(\mathscr{G},\mathbf{y}\right)$, we define

- $R\left(\mathscr{G},\mathbf{y}\right)$ as the total number of edges in $\Gamma\left(\mathscr{G},\mathbf{y}\right)$,
- $T\left(\mathscr{G},\mathbf{y}\right)$ as the number of connected components in $\Gamma\left(\mathscr{G},\mathbf{y}\right)$ and
- $U\left(\mathscr{G},\mathbf{y}\right)$ as the size of the largest connected component of $\Gamma\left(\mathscr{G},\mathbf{y}\right)$.

And to sum up the above, the set of summary statistics that where registered in the reference tables for each simulated field $\mathbf{y}$ is

$$\mathbf{S}(\mathbf{y}) = \left\{R\left(\mathscr{G}_4,\mathbf{y}\right); R\left(\mathscr{G}_8,\mathbf{y}\right); T\left(\mathscr{G}_4,\mathbf{y}\right); T\left(\mathscr{G}_8,\mathbf{y}\right); U\left(\mathscr{G}_4,\mathbf{y}\right); U\left(\mathscr{G}_8,\mathbf{y}\right)\right\}$$

in the first and second experiments.

In the third experiment, the observed field $\mathbf{y}$ takes values in $\mathbb{R}$ and we cannot apply directly the approach based on induced graphes because no two pixels share the same color. All of the above statistics are meaningless, including the statistics $R\left(\mathscr{G},\mathbf{y}\right)$ used by Grelaud et al. (2009) in the noise-free case. We rely on a quantization preprocessing performed via a kmeans algorithm on the observed colors that forgets the spatial structure of the field. The algorithm was tuned to uncover the same number of groups of colors as the number of latent colors, namely $K = 2$. If $q_2(\mathbf{y})$ denotes the resulting field, the set of summary statistics becomes

$$\mathbf{S}(\mathbf{y}) = \Big\{R\left(\mathscr{G}_4, q_2(\mathbf{y})\right); R\left(\mathscr{G}_8, q_2(\mathbf{y})\right); T\left(\mathscr{G}_4, q_2(\mathbf{y})\right); T\left(\mathscr{G}_8, q_2(\mathbf{y})\right); U\left(\mathscr{G}_4, q_2(\mathbf{y})\right); U\left(\mathscr{G}_8, q_2(\mathbf{y})\right)\Big\}.$$

We have assumed here that the number of latent colors is known to keep the same purpose of selecting the correct neighborhood structure. Indeed Cucala and Marin

(2013) have already proposed a (complex) Bayesian method to infer the appropriate number of hidden colors. But more generally, we can add statistics based on various quantizations $q_k(\mathbf{y})$ of $\mathbf{y}$ with k groups.

3.4.3 Numerical results

In all three experiments, we compare three nested sets of summary statistics $\mathbf{S}_{2D}(\mathbf{y})$, $\mathbf{S}_{4D}(\mathbf{y})$ and $\mathbf{S}_{6D}(\mathbf{y})$ of dimension 2, 4 and 6 respectively. They are defined as the projection onto the first two (respectively four and six) axes of $\mathbf{S}(\mathbf{y})$ described in the previous section. We stress here that $\mathbf{S}_{2D}(\mathbf{y})$, which is composed of the summaries given by Grelaud et al. (2009), are used beyond the noise-free setting where they are sufficient for model choice. In order to study the information carried by the connected components, we add progressively our geometric summary statistics to the first set, beginning by the $T\big(\mathscr{G},\mathbf{y}\big)$-type of statistics in $\mathbf{S}_{4D}(\mathbf{y})$. Finally, remark that, before evaluating the Euclidean distance in ABC algorithms, we normalize the statistics in each reference tables with respect to an estimation of their standard deviation since all these summaries take values on axis of different scales. Simulated images have been drawn thanks to the Swendsen and Wang (1987) algorithm. In the least favourable experiment, simulations of one hundred pictures (on pixel grid of size 100×100) via $20,000$ iterations of this Markovian algorithm when parameters drawn from our prior requires about one hour of computation on a single CPU with our optimized `C++` code. Hence the amount of time required by ABC is dominated by the simulations of y via the Swedsen-Wang algorithm. This motivated Moores, Drovandi, Mengersen, and Robert (2015) to propose a cut down on the cost of running an ABC experiment by removing the simulation of an image from hidden Potts model, and replacing it by an approximate simulation of the summary statistics. Another alternative is the clever sampler of Mira et al. (2001) that provides exact simulations of Ising models and can be extended to Potts models.

First experiment. Fig. 3.2(a) illustrates the calibration of the number of nearest neighbors (parameter k of Algorithm 10) by showing the evolution of the prior error rates (evaluated on a validation reference table including $20,000$ simulations) when k increases. We compared the errors of six classifiers to inspect the differences between the three sets of summary statistics (in yellow, green and magenta) and the impact of the size of the training reference table ($100,000$ simulations in solid lines; $50,000$ simulations in dashed lines). The numerical results exhibit that a good calibration of k can reduce the prior misclassification error. Thus, without really degrading the performance of the classifiers, we can reduce the amount of simulations required in

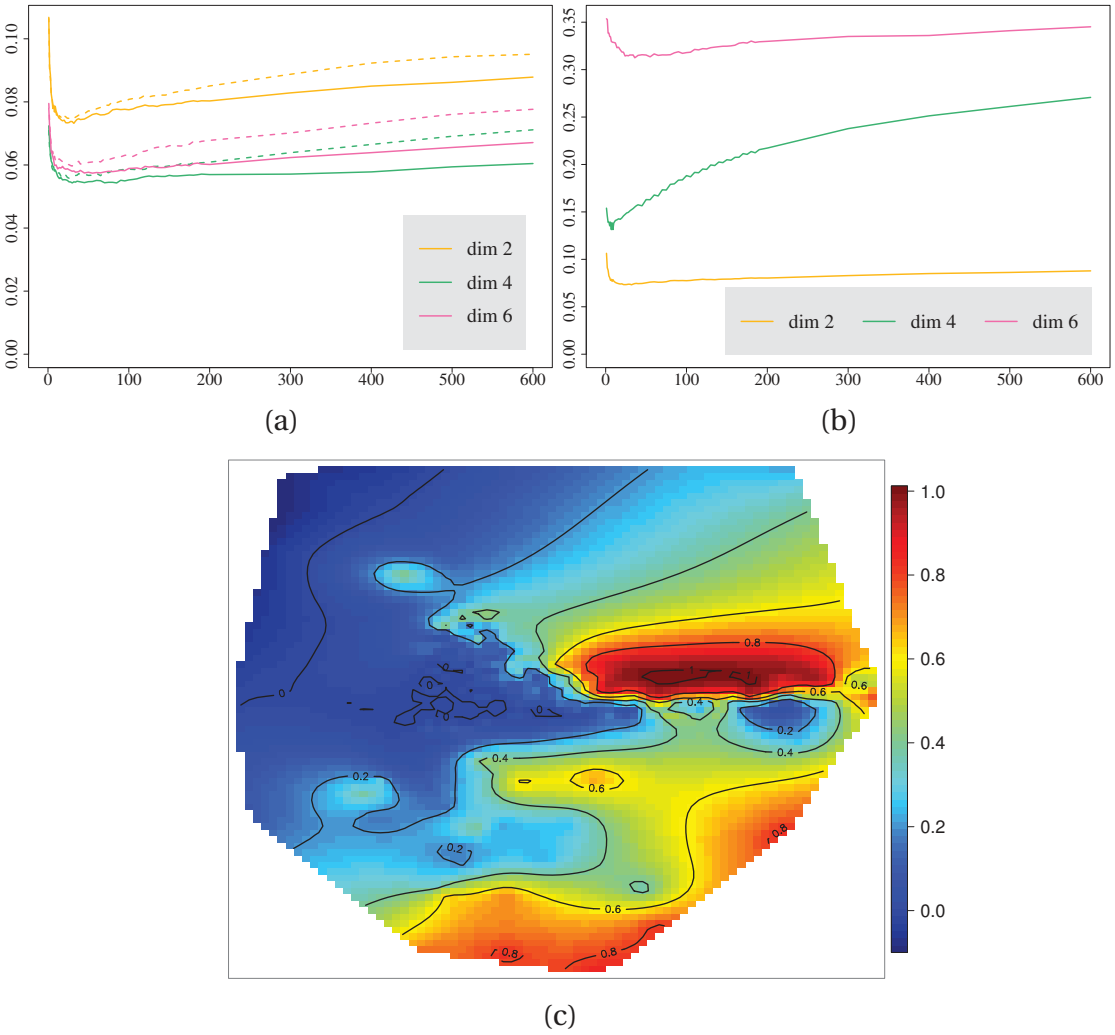


Figure 3.2: First experiment results. (a) Prior error rates (vertical axis) of ABC with respect to the number of nearest neighbors (horizontal axis) trained on a reference table of size 100,000 (solid lines) or 50,000 (dashed lines), based on the 2D, 4D and 6D summary statistics. (b) Prior error rates of ABC based on the 2D summary statistic compared with 4D and 6D summary statistics including additional ancillary statistics. (c) Evaluation of the local error on a 2D surface.

the training reference table, whose computation cost (in time) represents the main obstacle of ABC methods, see also Table 3.1. Moreover, as can be guessed from Fig. 3.2(a), the sizes of the largest connected components of induced graphs (included only in $\mathbf{S}_{6D}(\mathbf{y})$) do not carry additional information regarding the model choice and Table 3.1 confirms this results through evaluations of the errors on a test reference table of 30,000 simulations drawn independently of both training and validation reference tables.

Table 3.1: Evaluation of the prior error rate on a test reference table of size 30,000 in the first experiment.

Prior error rates		
Train size	**5,000**	**100,000**
2D statistics	8.8%	7.9%
4D statistics	6.5%	6.1%
6D statistics	7.1%	7.1%
Adaptive ABC	6.2%	5.5%

One can argue that the curse of dimensionality does not occur with such low dimensional statistics and sizes of the training set, but this intuition is wrong, as shown in Fig. 3.2(b). The latter plot shows indeed the prior misclassification rate as a function of k when we replace the last four summaries by ancillary statistics drawn independently of m and $\mathbf{y}$. We can conclude that, although the three sets of summary statistics carry then the same information in this artificial setting, the prior error rates increase substantially with the dimension (classifiers are not trained on infinite reference tables!). This conclusion shed new light on the results of Fig. 3.2(a): the $U\left(\mathscr{G},\mathbf{y}\right)$-type summaries, based on the size of the largest component, are not concretely able to help discriminate among models, but are either highly correlated with the first four statistics; or the resolution (in terms of size of the training reference table) does not permit the exploitation of the possible information they add.

Fig. 3.2(c) displays the local error rate with respect to a projection of the image space on a plan. We have taken here $\mathbf{S}_1(\mathbf{y}) = \mathbf{S}_{2D}(\mathbf{y})$ in Definition 5. And $\mathbf{S}_2(\mathbf{y})$ ranges a plan given by a projection of the full set of summaries that has been tuned empirically in order to gather the errors committed by calls of $\widehat{m}(\mathbf{S}_{2D}(\mathbf{y}))$ on the validation reference table. The most striking fact is that the local error rises above 0.9 in the oval, reddish area of Fig. 3.2(c). Other reddish areas of Fig. 3.2(c) in the bottom of the plot correspond to parts of the space with very low probability, and may be a dubious extrapolation of the Kriging algorithm. We can thus conclude that the information of the new geometric summaries depends highly on the position of y in the image space and have confidence in the interest of Algorithm 12 (adaptive ABC) in this framework. As exhibited in Table 3.1(d), this last classifier does not decrease dramatically the prior misclassification rates. But the errors of the non-adaptive classifiers are already low and the error of any classifier is bounded from below, as explained in Proposition 3.3. Interestingly though, the adaptive classifier relies on $\widehat{m}(\mathbf{S}_{2D}(\mathbf{y}))$ (instead of the most informative $\widehat{m}(\mathbf{S}_{6D}(\mathbf{y}))$) to take the final decision at about 60% of the images of our test reference table of size 30,000.

Table 3.2: Evaluation of the prior error rate on a test reference table of size 20,000 in the second experiment.

Prior error rates		
Train size	**50,000**	**100,000**
2D statistics	4.5%	4.4%
4D statistics	4.6%	4.1%
6D statistics	4.6%	4.3%

Second experiment. The framework was designed here to study the limitations of our approach based on the connected components of induced graphs. The number of latent colors is indeed relatively high and the noise process do not rely on any ordering of the colors to perturbate the pixels. Table 3.2 indicates the difficulty of capturing relevant information with the geometric summaries we propose. Only the sharpness introduced by a training reference table composed of 100,000 simulations distinguishes $\widehat{m}(\mathbf{S}_{4D}(\mathbf{y}))$ and $\widehat{m}(\mathbf{S}_{6D}(\mathbf{y}))$ from the basic classifier $\widehat{m}(\S_{2D}(\mathbf{y}))$. This conclusion is reinforced by the low value of number of neighbors after the calibration process, namely $k = 16$, 5 and 5 for $\widehat{m}(\mathbf{S}_{2D}(\mathbf{y}))$, $\widehat{m}(\mathbf{S}_{4D}(\mathbf{y}))$ and $\widehat{m}(\mathbf{S}_{6D}(\mathbf{y}))$ respectively. Hence we do not display in this Chapter other diagnosis plots based on the prior error rates or the conditional error rates, which led us to the same conclusion. The adaptive ABC algorithm did not improve any of these results.

Third experiment. The framework here includes a continuous noise process as described at the end of Section 3.4.1. We reproduced the entire diagnosis process performed in the first experiment and we obtained the results given in Fig. 3.3 and Table 3.3. The most noticeable difference is the extra information carried by the $U(\mathscr{G},\mathbf{y})$-statistics, representing the size of the largest connected component, and the adaptive ABC relies on the simplest $\widehat{m}(\mathbf{S}_{2D}(\mathbf{y}))$ in about 30% of the data space (measured with the prior marginal distribution in $\mathbf{y}$). Likewise, the gain in misclassification errors is not spectacular, albeit positive.

Table 3.3: Evaluation of the prior error rate on a test reference table of size 30,000 in the third experiment.

Prior error rates		
Train size	**5,000**	**100,000**
2D statistics	14.2%	13.8%
4D statistics	10.8%	9.8%
6D statistics	8.6%	6.9%
Adaptive ABC	8.2%	6.7%

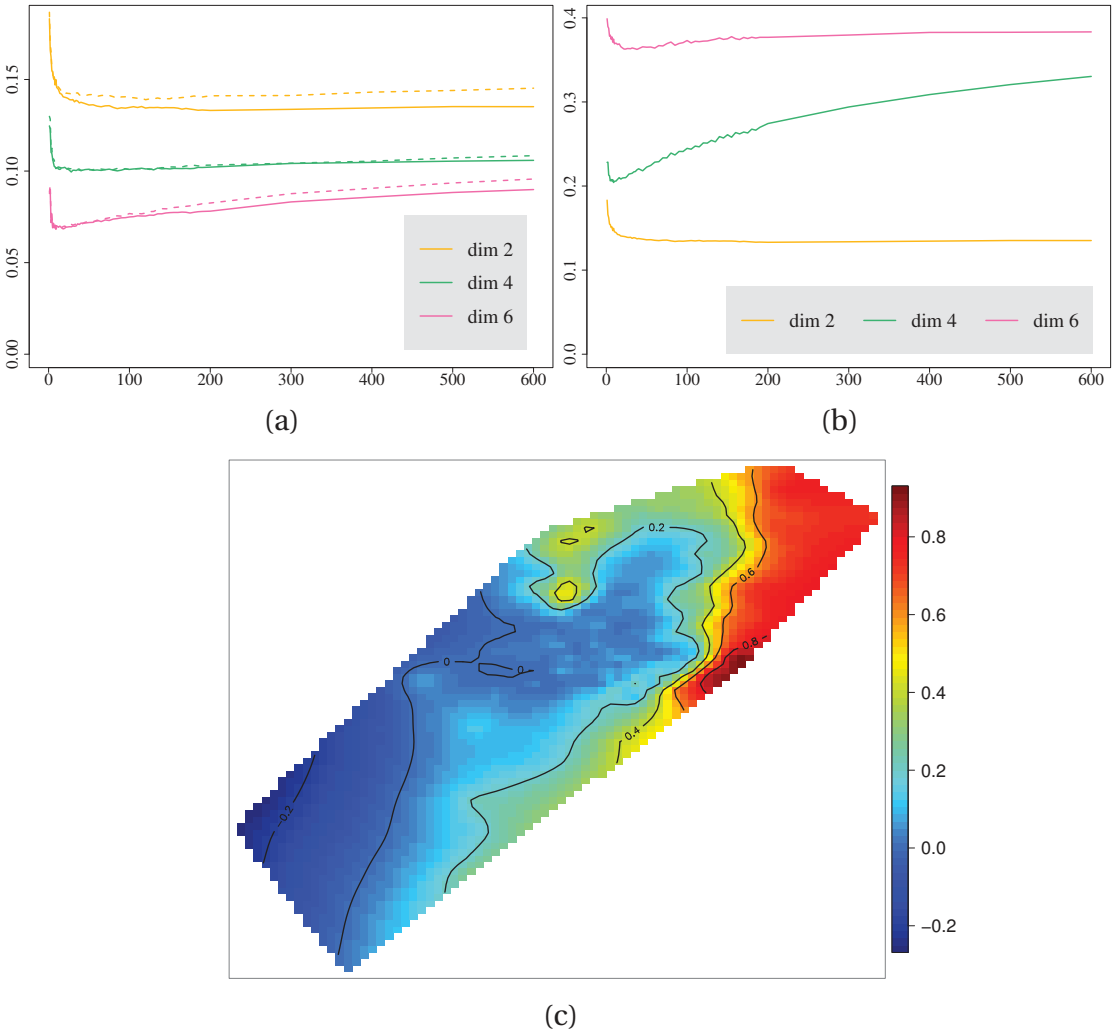


Figure 3.3: Third experiment results. (a) Prior error rates (vertical axis) of ABC with respect to the number of nearest neighbors (horizontal axis) trained on a reference table of size 100,000 (solid lines) or 50,000 (dashed lines), based on the 2D, 4D and 6D summary statistics. (b) Prior error rates of ABC based on the 2D summary statistics compared with 4D and 6D summary statistics including additional ancillary statistics. (c) Evaluation of the local error on a 2D surface.

4 Model choice criteria for hidden Gibbs random field

Throughout Chapter 3, the number of latent states K was assumed to be known but in many concrete situation this is not the case. Shaped by the development of Geman and Geman (1984) and Besag (1986), hidden Markov random fields have enjoyed great success in image segmentation, where one aims at estimating an unknown class assignment from the observation of a noisy copy of a latent random field. In this Chapter, we are interested in the joint selection of the dependency structure and the number of latent states of a hidden Potts model. The Bayesian Information Criterion (BIC, Schwarz, 1978) is an asymptotical estimate of the evidence that allows to answer the question from a Bayesian viewpoint but its exact computation within the context of hidden Markov random field is not feasible due to the intrinsic challenges of intractable likelihoods. Up to our knowledge, few attention has been paid to this specific issue in the literature aside from the work of Stanford and Raftery (2002) and Forbes and Peyrard (2003). The solution of Forbes and Peyrard (2003) derives from mean field theory which approximates the Markov random field by a system of independent variables whose distributions are fully tractable. This approach has proven great efficiency as regards to parameter estimation (Celeux et al., 2003) but surprisingly encounters difficulties to select a number of latent states. We could have also mentioned the suggestion of Cucala and Marin (2013) but from an Integrated Completed Likelihood estimation point of view. Nevertheless their complex algorithm is time consuming and cannot be easily extended to wider scope such as the choice of a dependency graph.

In this chapter, we propose approximations of BIC. The general approach described in Section 4.1 is to replace the intractable Gibbs distribution with a system of independent random vectors, namely blocks of the lattice, by taking advantage of the recursive algorithm of Section 1.3. To illustrate the performance of our approximations of BIC, in Section 4.3 we focus on three different experiments on simulated data. In particular

we address the question of the selection of the number of colors and the neighborhood system of a hidden Potts model. We conclude the numerical part of this Chapter with a comparison between the ABC procedures of Chapter 3 and our model choice criterion.

4.1 Block Likelihood Information Criterion

4.1.1 Background on Bayesian Information Criterion

The Bayesian Information Criterion offers a mean arising from Bayesian viewpoint to select a statistical model. This Section is a brief reminder on the construction of BIC and we refer the reader for instance to Raftery (1995) for a more detailed presentation.

We are given n independent and identically distributed observations $\mathbf{y} = \{y_1, \ldots, y_n\}$ from an unknown statistical model to estimate. The Bayesian approach to model selection is based on posterior model probabilities. Consider a finite set of models $\{m : 1, \ldots, M\}$ where each one is defined by a probability density function π_m related to a parameter space Θ_m. The model that best fits an observation $\mathbf{y}$ is the model with the highest posterior probability

$$\pi\left(m \,\middle|\, \mathbf{y}\right) = \frac{\pi(m)\, e\left(\mathbf{y} \,\middle|\, m\right)}{\sum_{m'} \pi(m')\, e\left(\mathbf{y} \,\middle|\, m'\right)},$$

where $e\left(\mathbf{y} \,\middle|\, m\right)$ denotes the evidence of m, that is the joint distribution of $(\mathbf{y}, \theta_m)$ integrated over space parameter Θ_m

$$e\left(\mathbf{y} \,\middle|\, m\right) = \int \pi_m\left(\mathbf{y} \,\middle|\, \theta_m\right) \pi_m\left(\theta_m\right) \mathrm{d}\theta_m.$$

Under the assumption of model being equally likely *a priori*, it is equivalent to choose the model with the largest evidence.

BIC is an asymptotical estimate of the evidence based on Laplace method for integral (*e.g.*, Schwarz, 1978, Tierney and Kadane, 1986, Raftery, 1995) defined by

$$-2\log e\left(\mathbf{y} \,\middle|\, m\right) \simeq \mathrm{BIC}(m) = -2\log \pi_m\left(\mathbf{y} \,\middle|\, \hat{\theta}_{\mathrm{MLE}}\right) + d_m \log(n), \tag{4.1}$$

where $\hat{\theta}_{\mathrm{MLE}}$ is the maximum likelihood estimator of π_m and d_m is the number of free parameters for model m. The $d_m \log(n)$ term corresponds to a penalty term which increases with the complexity of the model. Thus selecting the model with the largest evidence is equivalent to choose the model which minimizes BIC.

The definition of BIC derives from the following result for which proof can be found, amongst others, in the orignal paper of Schwarz (1978). We refer also the reader to Tierney and Kadane (1986) for a formalization of some of the arguments.

Proposition 4.2. *Under regularity conditions, the evidence of model m can be written as*

$$\log e\left(\mathbf{y} \mid m\right) = \log \pi_m\left(\mathbf{y} \mid \hat{\theta}_{MLE}\right) - \frac{d_m}{2}\log(n) + R_m\left(\hat{\theta}_{MLE}\right) + \mathcal{O}\left(n^{-\frac{1}{2}}\right), \tag{4.2}$$

where R_m is bounded as the sample size grows to infinity.

Proof. Consider the Taylor series expansion of $g_m(\theta) = \log\left\{\pi_m(\mathbf{y} \mid \theta)\pi_m(\theta_m)\right\}$ about the posterior mode, that is about the value θ^* that maximizes $g(\theta)$. The expansion writes as

$$g_m(\theta) = g_m\left(\theta^*\right) + \frac{1}{2}\left(\theta - \theta^*\right)^T \nabla^2 g_m\left(\theta^*\right)\left(\theta - \theta^*\right) + o\left(\left\|\theta - \theta^*\right\|\right).$$

It follows from the Laplace method for integrals applied to $e\left(\mathbf{y} \mid m\right) = \int \exp\left\{g_m(\theta)\right) \mathrm{d}\theta$ that

$$e\left(\mathbf{y} \mid m\right) = \log \pi_m\left(\mathbf{y} \mid \theta^*\right) + \log \pi_m\left(\theta^*\right) + \frac{d_m}{2}\log(2\pi) - \frac{1}{2}\det\left(A_m\right) + \mathcal{O}\left(n^{-1}\right), \tag{4.3}$$

where $A_m = -\nabla^2 g_m\left(\theta^*\right)$.

The idea now is that in large samples θ^* can be approximated with the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ and the Hessian matrix A_m with n times the Fisher information matrix $\mathbf{I}\left(\hat{\theta}_{\mathrm{MLE}}\right)$ for one observation, that is $A_m = n\mathbf{I}\left(\hat{\theta}_{\mathrm{MLE}}\right)$ where $\mathbf{I}\left(\hat{\theta}_{\mathrm{MLE}}\right) = -\mathbf{E}\left\{\nabla^2 \log \pi_m\left(Y_1 \mid \hat{\theta}_{\mathrm{MLE}}\right)\right\}$, the expectation being taken with respect to the density of Y_1. Hence the determinant of the Hessian matrix taken at the maximum likelihood estimator can be estimated by $\det(A_m) \approx n^{d_m}\det\left(\mathbf{I}\left(\hat{\theta}_{\mathrm{MLE}}\right)\right)$. These two approximations lead to an $\mathcal{O}\left(n^{-\frac{1}{2}}\right)$ error into equation (4.3).

The result immediately follows with

$$R_m\left(\hat{\theta}_{\mathrm{MLE}}\right) = \log \pi_m\left(\theta^*\right) + \frac{d_m}{2}\log(2\pi) - \frac{1}{2}\det\left(\mathbf{I}\left(\hat{\theta}_{\mathrm{MLE}}\right)\right). \qquad \square$$

Proposition 4.2 means that regardless of the prior on parameter, the error is, in general, solely bounded and does not go to zero even with an infinite amount of data. The approximation may hence seem somewhat crude. 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. In addition, a reasonable choice of the prior can lead to much smaller error. Indeed, Kass and Wasserman (1995) have found that the error is $\mathscr{O}\left(n^{-1/2}\right)$ for a well chosen multivariate normal prior distribution.

BIC can be defined beside the special case of independent random variables. In the latter case the approximations at the core of the proof of Proposition 4.2 are slightly different and the number of free parameter is, in general, not equal to the dimension of the parameter space as for the independent case. The consistency of BIC has been proven in various situations such as independent and identically distributed processes from the exponential families (Haughton, 1988), mixture models (Keribin, 2000), Markov chains (Csiszár et al., 2000, Gassiat, 2002) and Markov random fields for the selection of a neighborhood system (Csiszár and Talata, 2006).

When dealing with Markov random fields, penalized likelihood criteria like BIC faces the problem of an intractable likelihood. In addition to this issue that has been widely underlined in this dissertation, the number of free parameters in the penalty term has no simple formula. In the context of selecting a neighborhood system, Csiszár and Talata (2006) proposed to replace the likelihood by the pseudolikelihood and modify the penalty term as the number of all possible configurations for the neighboring sites. The resulting criterion is shown to be consistent as regards this model choice.

Up to our knowledge such a result has not been yet derived for hidden Markov random field for which another challenge appears as the incomplete likelihood requires to integrate over the latent

$$\pi_m\left(\mathbf{y} \,\middle|\, \theta\right) = \int_{\mathscr{X}} \pi\left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi\right) \pi\left(\mathbf{x} \,\middle|\, \beta, \mathscr{G}\right) \mu(\mathrm{d}\mathbf{x}). \tag{4.4}$$

The problem of approximating BIC could be once again termed a triple intractable problem since neither the maximum likelihood estimate $\hat{\theta}_{\mathrm{MLE}}$ nor the incomplete likelihood $\pi_m(\cdot \mid \theta)$ can be computed with standard methods and no simple definition of d_m is available.

Newton and Raftery (1994) tackled the issue of approximating BIC with an importance sampling procedure to supersede the crude Monte Carlo estimate based on computation over all realisations of a Markov chain with stationary distribution $\pi(\cdot \mid \psi, \mathscr{G})$. Such an approach could have been set up here by considering the likelihood of the conditional field $\pi(\cdot \mid \mathbf{y}, \theta, \mathscr{G})$ as an importance sampling function but follows time consuming procedure that we have decided not to pursue. In what follows, we rather focus on approximations of the criterion based on approximations of the Gibbs distribution such as the pseudolikelihood (Stanford and Raftery, 2002) or the mean

field-like approximations (Forbes and Peyrard, 2003).

4.2.1 Gibbs distribution approximations

The alternate proposal to simulation methods is to replace the Gibbs distribution by an approximation. As for the pseudolikelihood (Besag, 1975), the main idea consists in replacing the original Markov distribution by a product easier to deal with. But while pseudolikelihood is not a genuine probability distribution for Gibbs random field, the focus hereafter is solely on valid probability function by considering system of independent variables. This choice is motivated by the observations of Chapter 2 that at finite sample size, misspecification of the model has to be taken into account, so that constant terms may appear in the remainder R_m of Proposition 4.2.

Finding good approximations of the Gibbs distribution has long standing antecedents in statistical mechanics when one aims at predicting the response to the system to a change in the Hamiltonian. One important technique is based on a variational approach as the minimizer of the free energy, sometimes referred to as variational or Gibbs free energy, defined with the Kullback-Leibler divergence between $\mathbf{P}$ and the target distribution $\pi(\cdot \mid \psi, \mathscr{G})$ as

$$F(\mathbf{P}) = -\log Z\left(\psi, \mathscr{G}\right) + \mathrm{KL}\left(\mathbf{P}, \pi(\cdot \mid \psi, \mathscr{G})\right). \tag{4.5}$$

The Kullback-Leibler divergence being non-negative and zero if and only if $\mathbf{P} = \pi(\cdot \mid \psi, \mathscr{G})$, the free energy has an optimal lower bound achieved for $\mathbf{P} = \pi(\cdot \mid \psi, \mathscr{G})$. Minimizing the free energy with respect to the set of probability distribution on $\mathscr{X}$ allows to recover the Gibbs distribution but presents the same computational intractability. A solution is to minimize the Kullback-Leibler divergence over a restricted class of tractable probability distribution on $\mathscr{X}$. This is the basis of mean field approaches that aim at minimizing the Kullback-Leibler divergence over the set of probability functions that factorize on sites of the lattice (see also Section 1.8.3). The minimization of (4.5) over this set leads to fixed point equations for each marginal of $\mathbf{P}$ (see for example Jordan et al., 1999).

Instead of considering distributions that completely factorize on single sites, we are interested in tractable approximations that factorize over larger sets of nodes, namely blocks of the lattice. Consider a partition of $\mathscr{S}$ into contiguous rectangular blocks, namely

$$\mathscr{S} = \bigsqcup_{\ell=1}^{C} A(\ell),$$

and denote $\tilde{D}$ the class of independent probability distributions $\mathbf{P}$ that factorize with respect to this partition, that is if $\mathscr{X}_{A(\ell)} = \prod_{i \in A(\ell)} \mathscr{X}_i$ stands for the configuration space of the block $A(\ell)$, for all $\mathbf{x}$ in $\mathscr{X}$

$$\mathbf{P}(\mathbf{x}) = \prod_{\ell=1}^{C} \mathbf{P}_\ell \left(x_{A(\ell)} \right), \text{ where } \mathbf{P}_\ell \in \mathscr{M}_1^+ \left(\mathscr{X}_{A(\ell)} \right) \text{ and } \mathbf{P} \in \mathscr{M}_1^+ (\mathscr{X}).$$

Our proposal is to derive BIC approximations by replacing the intractable Gibbs distribution with a well chosen probability distribution in $\tilde{D}$.

To take over from the Gibbs likelihood, we propose probability distribution of the form

$$\mathbf{P}\left(\mathbf{x} \,\middle|\, \tilde{\mathbf{x}}, A(1), \ldots, A(C), \psi \right) = \prod_{\ell=1}^{C} \pi \left(\mathbf{x}_{A(\ell)} \,\middle|\, \mathbf{X}_{B(\ell)} = \tilde{\mathbf{x}}_{B(\ell)}, \psi, \mathscr{G} \right), \tag{4.6}$$

where $\tilde{\mathbf{x}}$ is a constant field in $\mathscr{X}$ to specify and $B(\ell)$ is either the border of $A(\ell)$, *i.e.*, elements of the absolute complement of $A(\ell)$ that are connected to elements of $A(\ell)$ in $\mathscr{G}$, or the empty set. In the latter case, we are cancelling the edges in $\mathscr{G}$ that link elements of $A(\ell)$ to elements of any other subset of $\mathscr{S}$ such that the factorization is independent of $\tilde{\mathbf{x}}$. The Gibbs distribution is simply replaced by the product of the likelihood restricted to $A(\ell)$. For instance a Potts model on $\mathscr{X}$ is replaced with a product of Potts models on $\mathscr{X}_{A(\ell)}$. To underline that point, $\tilde{\mathbf{x}}$ is omitted in what follows when $B(\ell) = \emptyset$. Note that the composite likelihood (2.1) differs from (4.6) in most cases since blocks are not allowed to overlap and contrary to conditional composite likelihoods, neighbors are set to constants. The only example of composite likelihoods that lies in $\tilde{D}$ is marginal composite likelihoods for non overlapping blocks.

The assumption of independent blocks leads to tractable BIC approximations. Indeed, plugging the probability distribution (4.6) in place of the Gibbs distribution in (4.4) yields

$$\begin{aligned}
\mathbf{P}_m\left(\mathbf{y} \,\middle|\, \tilde{\mathbf{x}}, \theta \right) &= \sum_{\mathbf{x} \in \mathscr{X}} \pi \left(\mathbf{y} \,\middle|\, \mathbf{x}, \phi \right) \mathbf{P}\left(\mathbf{x} \,\middle|\, \tilde{\mathbf{x}}, A(1), \ldots, A(C), \psi \right) \\
&= \prod_{\ell=1}^{C} \sum_{\mathbf{x}_{A(\ell)}} \left\{ \prod_{i \in A(\ell)} \pi \left(y_i \,\middle|\, x_i, \phi \right) \right\} \pi \left(\mathbf{x}_{A(\ell)} \,\middle|\, \mathbf{X}_{B(\ell)} = \tilde{\mathbf{x}}_{B(\ell)}, \psi, \mathscr{G} \right) \\
&= \prod_{\ell=1}^{C} \sum_{\mathbf{x}_{A(\ell)}} \pi \left(\mathbf{y}_{A(\ell)} \,\middle|\, \mathbf{x}_{A(\ell)}, \phi \right) \pi \left(\mathbf{x}_{A(\ell)} \,\middle|\, \mathbf{X}_{B(\ell)} = \tilde{\mathbf{x}}_{B(\ell)}, \psi, \mathscr{G} \right).
\end{aligned} \tag{4.7}$$

This estimate of the incomplete likelihood $\pi_m(\cdot \mid \theta)$ leads to the following BIC approxi-

mations

$$\mathrm{BIC}(m) \approx -2\log \mathbf{P}_m\left(\mathbf{y} \,\middle|\, \tilde{\mathbf{x}}, \theta^*\right) + d_m \log(|\mathscr{S}|) := \mathrm{BLIC}^{\tilde{\mathbf{x}}}\left(m \,\middle|\, \theta^*\right), \tag{4.8}$$

where $\theta^* = \left(\phi^*, \psi^*\right)$ is a parameter value to specify. We refer to these approximations as Block Likelihood Information Criterion (BLIC). In the first instance, the number of free parameters d_m is set to the dimension of Θ_m, that is we are neglecting the interaction between variables within a block in the penalty term.

Our proposal relies on that each term of the product (4.7) can be computed using the recursion of Friel and Rue (2007) as long as the blocks are small enough (see Section 1.3). Indeed for models whose potential linearly depends on the parameter, that is $H(\mathbf{x} \,|\, \psi, \mathscr{G}) = \psi^T \mathbf{S}(\mathbf{x})$ with $\mathbf{S}$ a vector of sufficient statistics, the probability distribution on $A(\ell)$ can be written as a Gibbs distribution on the block conditioned on the fixed border $\tilde{\mathbf{x}}_{B(\ell)}$, namely

$$\pi\left(\mathbf{x}_{A(\ell)} \,\middle|\, \mathbf{X}_{B(\ell)} = \tilde{\mathbf{x}}_{B(\ell)}, \psi, \mathscr{G}\right) = \frac{1}{Z\left(\psi, \mathscr{G}, \tilde{\mathbf{x}}_{B(\ell)}\right)} \exp\left(\psi^T \mathbf{S}\left(\mathbf{x}_{A(\ell)} \,\middle|\, \tilde{\mathbf{x}}\right)\right),$$

where $\mathbf{S}\left(\mathbf{x}_{A(\ell)} \,\middle|\, \tilde{x}\right)$ is the restriction of $\mathbf{S}$ to the subgraph defined on the set $A(\ell)$ and conditioned on the fixed border $\tilde{\mathbf{x}}_{B(\ell)}$ (see Example 2.1.1). Assuming that all the marginals of the emission distribution are positive, it follows

$$\begin{aligned}
&\sum_{\mathbf{x}_{A(\ell)}} \pi\left(\mathbf{y}_{A(\ell)} \,\middle|\, \mathbf{x}_{A(\ell)}, \phi\right) \pi\left(\mathbf{x}_{A(\ell)} \,\middle|\, \mathbf{X}_{B(\ell)} = \tilde{\mathbf{x}}_{B(\ell)}, \psi, \mathscr{G}\right) \\
&\qquad = \frac{1}{Z\left(\psi, \mathscr{G}, \tilde{\mathbf{x}}_{B(\ell)}\right)} \underbrace{\sum_{\mathbf{x}_{A(\ell)}} \exp\left\{\log \pi\left(\mathbf{y}_{A(\ell)} \,\middle|\, \mathbf{x}_{A(\ell)}, \phi\right) + \psi^T \mathbf{S}\left(\mathbf{x}_{A(\ell)} \,\middle|\, \tilde{\mathbf{x}}\right)\right\}}_{= Z\left(\theta, \mathscr{G}, \mathbf{y}_{A(\ell)}, \tilde{\mathbf{x}}_{B(\ell)}\right)}.
\end{aligned}$$

The term $Z\left(\theta, \mathscr{G}, \mathbf{y}_{A(\ell)}, \tilde{\mathbf{x}}_{B(\ell)}\right)$ corresponds to the normalizing constant of the conditional random field $\mathbf{X}_{A(\ell)}$ knowing $\mathbf{Y}_{A(\ell)} = \mathbf{y}_{A(\ell)}$ and $\mathbf{X}_{B(\ell)} = \tilde{\mathbf{x}}_{B(\ell)}$, that is the initial model with an extra potential on singletons. Then the algebraic simplification at the core of Algorithm 3 applies for both normalizing constants, such that we can exactly compute the Block Likelihood Information Criterion, namely

$$\begin{aligned}
\mathrm{BLIC}^{\tilde{\mathbf{x}}}\left(m \,\middle|\, \theta^*\right) = -2 \sum_{\ell=1}^{C} \Big\{ &\log Z\left(\theta^*, \mathscr{G}, \mathbf{y}_{A(\ell)}, \tilde{\mathbf{x}}_{B(\ell)}\right) \\
&- \log Z\left(\psi^*, \mathscr{G}, \tilde{\mathbf{x}}_{B(\ell)}\right) \Big\} + d_m \log(|\mathscr{S}|).
\end{aligned} \tag{4.9}$$

4.2.2 Related model choice criteria

This approach encompasses the Pseudolikelihood Information Criterion (PLIC, (1.38)) of Stanford and Raftery (2002) as well as the mean field-like approximations $\mathrm{BIC}^{\text{MF-like}}$ (1.37) proposed by Forbes and Peyrard (2003) (see Section 1.8.3). When one considers the finest partition of $\mathscr{S}$, that is distributions that factorize on sites, they have already proposed ingenious solutions for choosing $\tilde{\mathbf{x}}$ and estimating $\hat{\theta}_{\mathrm{MLE}}$ in (4.8). Indeed, Stanford and Raftery (2002) suggest to set $(\tilde{\mathbf{x}}, \hat{\theta}_{\mathrm{MLE}})$ to the final estimates $(\hat{\theta}^{\mathrm{ICM}}, \tilde{\mathbf{x}}^{\mathrm{ICM}})$ of the unsupervised Iterated Conditional Modes (ICM, Besag, 1986) algorithm, while Forbes and Peyrard (2003) put forward the use of the output $(\hat{\theta}^{\text{MF-like}}, \tilde{\mathbf{x}}^{\text{MF-like}})$ of the simulated field algorithm of Celeux et al. (2003) (see Algorithm 5 in Section 1.5.2). To make this statement clear, we could note

$$\begin{aligned}\mathrm{PLIC}(m) &= \mathrm{BLIC}^{\tilde{\mathbf{x}}^{\mathrm{ICM}}}\left(m \,\middle|\, \hat{\theta}^{\mathrm{ICM}}\right),\\ \mathrm{BIC}^{\text{MF-like}}(m) &= \mathrm{BLIC}^{\tilde{\mathbf{x}}^{\text{MF-like}}}\left(m \,\middle|\, \hat{\theta}^{\text{MF-like}}\right).\end{aligned}$$

Whilst PLIC shows good result as regards the selection of the number of components of the hidden state, ICM performs poorly for the parameter estimation in comparison with the EM-like algorithm of Celeux et al. (2003). Hence we advocate in favour of the latter in what follows to get estimates of $\hat{\theta}_{\mathrm{MLE}}$ and to fix a segmented random field $\tilde{\mathbf{x}}$.

We shall also remark that for a factorization over the graph nodes when $B(\ell) = \emptyset$ we retrieve a mixture model. Indeed, turning off all the edges in $\mathscr{G}$ leads to approximate the Gibbs distribution by a multinomial distribution with event probabilities depending on the potential on singletons. Hence if marginal emission distribution are Gaussian random variables depending on the component on the latent site associated, we would deal with a classical Gaussian mixture model.

4.3 Comparison of BIC approximations

Our primary intent with the BIC approximations exposed in Section 4.1 was to choose the number of latent states and a dependency structure of a hidden Markov random fields. The following numerical experiments illustrate the performances as regards these questions for realizations of a hidden Potts model. Section 4.3.2 and Section 4.3.3 focus on a comparison between the different criteria to discriminate between hidden models in various settings while Section 4.3.4 presents a comparison with the ABC procedures introduced in Chapter 3.

4.3.1 Hidden Potts models

This numerical part of the Chapter focuses on observations for which the hidden field is modelled by a K-states Potts model parametrized by $\psi \equiv \beta$. The model sets a probability distribution on $\mathscr{X} = \{1,\dots,K\}^n$ defined by

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

We set the emission law $\pi(\mathbf{y} \mid \mathbf{x}, \phi) = \prod_{i \in \mathscr{S}} \pi(y_i \mid x_i, \phi)$, such that the marginal distribution are Gaussian distribution depending on the related latent state, namely

$$y_i \mid x_i = k \;\sim \mathscr{N}\left(\mu_k, \sigma_k^2\right) \quad k \in \{0,\dots,K-1\},$$

where μ_k is the mean and σ_k is the standard deviation for sites belonging to class k. The parameter to be estimated with the ICM or simulated field algorithms is then

$$\theta = \left(\phi, \beta\right), \text{ with } \phi = \left\{\left(\mu_k, \sigma_k\right) :\; k = 0,\dots,K-1\right\}.$$

We denote HPM$(\mathscr{G}, \theta, K)$, the hidden K-states Potts model defined above.

The common point of our examples is to select the hidden Potts model that better fits a given observation $\mathbf{y}^{\text{obs}}$ composed of $n = 100 \times 100$ pixels among a collection

$$\mathscr{M} = \left\{\text{HPM}\left(\mathscr{G}, \theta, K\right) :\; K = K_{\min},\dots,K_{\max}\,;\; \mathscr{G} \in \left\{\mathscr{G}_4, \mathscr{G}_8\right\}\right\},$$

where K is the number of colors of the corresponding model and $\mathscr{G}$ is one of the two possible neighborhood systems: $\mathscr{G}_4$ and $\mathscr{G}_8$ defined in Chapter 1, see Figure 1.1. For each model HPM$(\mathscr{G}, \theta, K)$, the estimate $\hat{\theta}_{\text{MLE}}$ and the segmented field $\tilde{\mathbf{x}}$ were computed using SpaCEM^3 (see the Documentation on http://spacem3.gforge.inria.fr). The software allows the implementation of the unsupervised ICM algorithm as well as the simulated field algorithm and provides computation of PLIC, the mean field-like approximations BIC$^{\text{MF-like}}$ and BIC$^{\text{GBF}}$ (Equation (1.40)). The ICM and the EM-like algorithms were both initialized with a simple K-means procedure. The stopping criterion is then settled to a number of 200 iterations that is enough to ensure the convergence of the procedure.

In what follows, we restrict each $A(\ell)$ to be of the same dimension and in particular square block of dimension $b \times b$. For the sake of clarity the Block Likelihood Criterion is indexed by the dimension of the blocks, namely for a partition of square blocks of size $b \times b$ for which $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{\text{MF-like}}$ and $\hat{\theta}_{\text{MLE}} = \hat{\theta}^{\text{MF-like}}$, we note it $\text{BLIC}^{\text{MF-like}}_{b \times b}$. As already

mentioned, we then have $\text{BIC}^{\text{MF-like}} = \text{BLIC}^{\text{MF-like}}_{1\times 1}$. We recall that when $B(\ell) = \emptyset$, $\tilde{\mathbf{x}}$ is omitted in the previous notations, that is for a square blocks partition we note our criterion $\text{BLIC}_{b\times b}$. Then $\text{BLIC}_{1\times 1}$ is the BIC approximations corresponding to a finite independent mixture model. All criterion were tested on simulated images obtained using the Swendsen-Wang algorithm. We describe below the different experiments settings we have considered and the results we got.

4.3.2 First experiment: selection of the number of colors

We considered realizations from hidden Potts models with $K_T = 4$ colors. We carried out 100 simulations from the first order neighborhood structure $\mathscr{G}_4$ and 100 simulations from the second order neighborhood structure $\mathscr{G}_8$. In this experiment the dependency structure is assumed to be known and the aim is to recover the number K of colors of the latent configuration. The interaction parameter β was set close to the phase transition, namely $\beta = 1$ and $\beta = 0.4$ for $\mathscr{G}_4$ and $\mathscr{G}_8$ respectively. These values of the parameter ensure the images present homogeneous regions and then the observations exhibit some spatial structure. Such settings illustrate the advantage of taking into account spatial information of the model. Obviously, for values of β where the interaction is weaker, the benefit of the criterion that include the dependency structure of the model is not clear. The latter could even be misleading in comparison with BIC approximations for independent mixture models when β is close to zero. On the other side, when β is above the phase transition, the distribution on $\mathscr{X}$ becomes heavily multi-modal and there is almost solely one class represented in the image regardless the number of colors of the model.

The noise process is a homoscedastic Gaussian noise centered at the value of the related nodes, namely

$$y_i \mid x_i = k \sim \mathcal{N}\left(k, \sigma_k^2\right) \quad k \in \{0, \ldots, K-1\},$$

where $\sigma_k = 0.5$ for $k = 1, \ldots, K$. Even though the noise model is homoscedastic, we still index the standard deviation by k since we do not use the assumption of a constant variance in the estimation procedure, such that the number of parameters estimated is $d_m = 2 \times k + 1$.

The results obtained for the different criterion are reported in Table 4.1. For $b \geq 2$, $\text{BLIC}_{b\times b}$ outperform the different criterion even though PLIC and $\text{BLIC}_{1\times 1}$ provide good results. By contrast approximations based on mean field-like approximations, that is $\text{BIC}^{\text{MF-like}}$, BIC^{GBF} and $\text{BLIC}^{\text{MF-like}}_{2\times 2}$, perform poorly. These conclusions need nonetheless to be put into perspective. Figure 4.1(a) shows that the main issue encoun-

Table 4.1: Selected K in the first experiment for 100 realizations from HPM($\mathcal{G}_4, \theta, 4$) and 100 realizations from HPM($\mathcal{G}_8, \theta, 4$) using Pseudolikelihood Information Criterion (PLIC), mean field-like approximations ($\text{BIC}^{\text{MF-like}}$, BIC^{GBF}) and Block Likelihood Information Criterion (BLIC) for various sizes of blocks and border conditions.

	HPM($\mathcal{G}_4, \theta, 4$)							HPM($\mathcal{G}_8, \theta, 4$)						
K	**2**	**3**	**4**	**5**	**6**	**7**	**K**	**2**	**3**	**4**	**5**	**6**	**7**
PLIC	0	9	91	0	0	0	PLIC	0	7	93	0	0	0
$\text{BIC}^{\text{MF-like}}$	0	0	39	23	16	22	$\text{BIC}^{\text{MF-like}}$	0	0	43	18	19	20
BIC^{GBF}	0	0	39	25	18	18	BIC^{GBF}	0	0	52	20	19	9
$\text{BLIC}^{\text{MF-like}}_{2\times2}$	0	0	58	18	8	16	$\text{BLIC}^{\text{MF-like}}_{2\times2}$	0	0	52	14	17	17
$\text{BLIC}_{1\times1}$	0	0	97	1	2	0	$\text{BLIC}_{1\times1}$	0	3	90	1	4	2
$\text{BLIC}_{2\times2}$	0	0	100	0	0	0	$\text{BLIC}_{2\times2}$	0	1	99	0	0	0
							$\text{BLIC}_{4\times4}$	0	0	100	0	0	0

tered by these criterion is their inability to discriminate between the more complex models. Indeed these BIC approximations reach a plateau from $K = 4$, a problem that other criterion do not face. As an example, Figure 4.1(b) and Figure 4.1(c) represent boxplots of the difference between BIC values for HPM($\mathcal{G}_4, \theta, K$) as K is increasing for the 100 realizations, namely

$$\Delta(K \to K+1) = \text{BIC}\left(\text{HPM}(\mathcal{G}, \hat{\theta}_{\text{MLE}}, K+1)\right) - \text{BIC}\left(\text{HPM}(\mathcal{G}, \hat{\theta}_{\text{MLE}}, K)\right),$$

for $K = K_{\min}, \ldots, K_{\max}$. Hence, BIC approximations grow with K if $\Delta(K \to K+1) \geq 0$ and decrease otherwise. It appears that $\text{BLIC}_{2\times2}$ increases systematically from $K = 4$ whereas $\text{BIC}^{\text{MF-like}}$ tend to be constant, or even decreases, so that none minimum can be clearly identified. We do not provide the boxplots for $\text{BIC}^{\text{MF-like}}$ and BIC^{GBF} because they are significantly the same.

Finally these results illustrate in particular the importance of a well chosen segmented field $\tilde{\mathbf{x}}$. Indeed PLIC and $\text{BIC}^{\text{MF-like}}$ are both criterion of type $\text{BLIC}^{\tilde{\mathbf{x}}}_{1\times1}$ but their performances greatly differ on this example. As regards the selection of K, $\text{BLIC}_{b\times b}$ circumvent this question whilst performing better.

4.3.3 Second experiment: selection of the dependency structure

For this second experiments the setting was exactly the same than for the first experiment. The only difference is that as first instance the number of colors K_T is assumed to be known while the neighborhood system has to be chosen. To answer such a

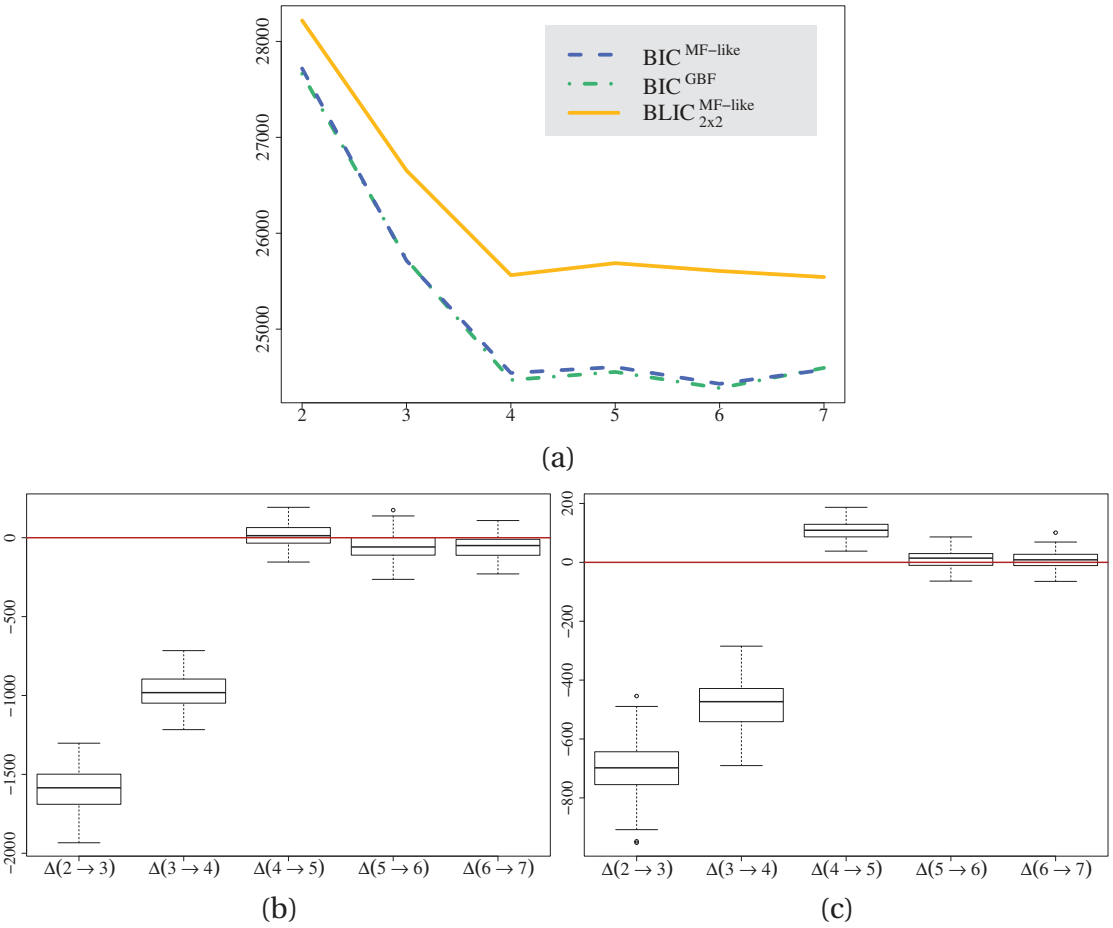


Figure 4.1: First experiment results. (a) $\text{BIC}^{\text{MF-like}}$, BIC^{GBF} and $\text{BLIC}^{\text{MF-like}}_{2\times2}$ values for one realization of a first order hidden Potts model HPM$(\mathcal{G},\theta,4)$. (b) Difference between $\text{BLIC}^{\text{MF-like}}_{2\times2}$ values for 100 realization of a first order hidden Potts model HPM$(\mathcal{G}_4,\theta,4)$ as K is increasing. (c) Difference between $\text{BLIC}_{2\times2}$ values for 100 realization of a first order hidden Potts model HPM$(\mathcal{G}_4,\theta,4)$ as K is increasing

question it is obvious that we can not use criterion $\text{BLIC}_{1\times1}$ based on independent mixture model.

As regards this question, all but two criterion perform very well, see Table 4.2. In the first place, PLIC faces trouble to select the correct $\mathcal{G}_4$. This illustrate the importance of the estimation of the interaction parameter β. We have observed that the ICM algorithm whilst providing good segmented field, produces poorer estimates of the parameter than the simulated field algorithm. This has an impact quite important since β sets the strength of interaction between neighboring nodes of the graph $\mathcal{G}$ and is most representative of the spatial correlation. On the other hand, $\text{BLIC}_{2\times2}$ fails to select the neighborhood system for second order hidden Potts model HPM$(\mathcal{G}_8,\theta,4)$. This conclusion can be simply explained by the fact that the block does not include

Table 4.2: Selected $\mathscr{G}$ in the second experiment for 100 realizations from HPM($\mathscr{G}_4, \theta, 4$) and 100 realizations from HPM($\mathscr{G}_8, \theta, 4$) using Pseudolikelihood Information Criterion (PLIC), mean field-like approximations ($\text{BIC}^{\text{MF-like}}$, BIC^{GBF}) and Block Likelihood Information Criterion (BLIC) for various sizes of blocks and border conditions.

HPM($\mathscr{G}_4, \theta, 4$)

	$\mathscr{G}_4$	$\mathscr{G}_8$
PLIC	53	47
$\text{BIC}^{\text{MF-like}}$	100	0
BIC^{GBF}	100	0
$\text{BLIC}^{\text{MF-like}}_{2\times2}$	100	0
$\text{BLIC}_{2\times2}$	100	0

HPM($\mathscr{G}_8, \theta, 4$)

	$\mathscr{G}_4$	$\mathscr{G}_8$
PLIC	0	100
$\text{BIC}^{\text{MF-like}}$	0	100
BIC^{GBF}	0	100
$\text{BLIC}^{\text{MF-like}}_{2\times2}$	0	100
$\text{BLIC}_{2\times2}$	59	41
$\text{BLIC}_{4\times4}$	0	100

enough spatial information to discriminate between the competing models. When the primary purpose is the selection of a dependency structure, we should use block large enough to be informative regarding the different neighborhood systems in competition.

Aside the two above exceptions, the good performances of all criteria can be surprising. The same experiment has been done for stronger noise with $\sigma_k = 0.75$ and $\sigma_k = 1$. The conclusion remains the same. It appears that for a conditionally independent noise process, neighborhood system are readily distinguished close to the phase transition. This is not true for any parameter value as illustrated in the third experiment.

In the second instance, we supposed that K_T and $\mathscr{G}$ were unknown, so that we were interested in the joint selection of the number of colors and of the dependency graph. For this example, the results remain the same than in Table 4.1 with the exception of PLIC. Indeed, the different criterion manage to differentiate the model in terms of the graph $\mathscr{G}$ so that their performances are directly related to their ability to choose the correct number of colors.

4.3.4 Third experiment: BLIC *versus* ABC

This third experiment is the occasion to compare BLIC with the ABC procedures introduced in Chapter 3. We return to the problem of solely selecting the dependency graph when the number of colors is know. We still consider a homoscedastic Gaussian noise whose marginal distribution is characterized by

$$y_i \mid x_i = k \sim \mathcal{N}(k, \sigma_k^2) \quad k \in \{0, 1\}$$

Table 4.3: Evaluation of the prior error rate of ABC procedures and of the error rate for the model choice criterion in the third experiment.

Train size	5,000	100,000	Criterion	Error rate
2D statistics	14.2%	13.8%	PLIC	19.8%
4D statistics	10.8%	9.8%	$\text{BIC}^{\text{MF-like}}$	7.6%
6D statistics	8.6%	6.9%	BIC^{GBF}	7.1%
Adaptive ABC	8.2%	6.7%	$\text{BLIC}_{4\times4}$	7.7%

but over bicolor Potts models. The standard deviation $\sigma_k = 0.39$, $k \in \{0,1\}$, was set so that the probability of a wrong prediction of the latent color with a marginal MAP rule on the Gaussian model is about 10% in the thresholding step of the ABC procedure. Regarding the dependency parameter β, we set prior distributions below the phase transition which occurs at different levels depending on the neighborhood structure. Precisely we used a uniform distribution over $(0;1)$ when the adjacency is given by $\mathcal{G}_4$ and a uniform distribution over $(0;0.35)$ with $\mathcal{G}_8$. In order to examine the performance of model choice criteria in comparison of ABC, we carried out 1000 realizations from $\text{HPM}(\mathcal{G}_4,\theta,2)$ and 1000 realizations from $\text{HPM}(\mathcal{G}_8,\theta,2)$ with parameters from the priors. The results are presented in Table 4.3

The novel ABC procedure introduced in Chapter 3 appears to provide the best performances but for a training reference table of size 100 000. This reinforces the idea that for unlimited computation possibilities, ABC can efficiently address situations where the likelihood is intractable. However, Table 4.3 suggest that for a much lower computational cost it is possible to get equivalent, or even better, error rate by using model choice criterion $\text{BIC}^{\text{MF-like}}$, BIC^{GBF} or $\text{BLIC}_{b\times b}$, while PLIC seems not to be overtaken. In this example, BIC^{GBF} slightly supersede $\text{BIC}^{\text{MF-like}}$ and $\text{BLIC}_{b\times b}$. This can be explained by the fact that for parameter from the prior close to zero, the assumption of independence between the sites is almost true. In the latter case, estimating BIC using the first order approximations of the partition function of Gibbs distribution (see the lower bound 1.39 in Section 1.8.3) may be preferable than using normalizing constants defined on blocks.

Conclusion

The present dissertation addresses well-known statistical inference issues for Markov random fields. One of our contribution concerns the calibration of approximate Bayesian computation algorithms for model choice as a classification problem. In this context, we have derived a local error rate that is an indicator of the gain or the loss of statistical information induced by the summary statistics conditional on the observed value. Consequently, we set up an adaptive classifier which is an attempt to fight locally against the curse of dimensionality. Our approach is an advance over most projection methods which are focused on parameter estimation (Blum et al., 2013). While most of these techniques perform a global trade off between the dimension and the information of the summary statistics over the whole prior domain, our proposal allows to select, at least in theory, the optimal set of summary statistic as the one minimizing the local error rate at the observed value. Principles of our proposal are well founded by avoiding the well-known optimism of the training error rates and by resorting to validation and test reference tables in order to evaluate the error practically. Another possibility taken by Pudlo et al. (2014) is to resort to a machine learning classifier adapted to a large number of covariates such as random forest.

Regarding latent Markov random fields, the proposed method of constructing summary statistics based on the induced graphs yields a promising route to construct relevant summary statistics in this framework. The approach is very intuitive and can be reproduced in other settings. For instance, if the goal is to select between isotropic latent Gibbs models and anisotropic models, the averaged ratio between the width and the length of the connected components or the ratio of the width and the length of the largest connected components can be relevant numerical summaries. We have also explained how to adapt the method to a continuous noise by performing a quantization of the observed values at each site of the fields. The detailed analysis of the numerical results demonstrates that the approach is promising. However the results on the 16 color example indicate the limitation of the induced graph approach as the number of colors grows. We believe that there exists a road we did not explore above with an induced graph that add weights on the edges of the graph according to

the proximity of the colors, but the grouping of sites on such weighted graph is not trivial.

The numerical results of Chapter 3 opens up new vistas regarding the calibration of ABC algorithms. Our contribution could be especially interesting for statistical models on high dimensional space for which ABC is the only solution available. Our work has put this statement into perspective for hidden Markov random fields. The numerical results in Chapter 3 highlighted that the calibration of the number of neighbors in ABC provides better results (in terms of misclassification) than a threshold set as a fixed quantile of the distances between the simulated and the observed datasets (as proposed in Marin et al., 2012). Consequently, we pointed out that the shift from posterior distribution regression to classification problem allows to reduce noticeably the number of simulation required in the ABC reference table without increasing the misclassification error rates. The latter represents an important conclusion since the simulation of a latent Markov random field requires a non-negligible amount of time. Even though the computational cost is cut down, ABC procedures remain time consuming in comparison with methods based on approximations of the likelihood and their relative efficiency in certain situations regarding Markov random fields can barely justify the extra computational cost. However, mention that we run into that problem because the computation of the summary statistics require to simulate from the model. Following Moores et al. (2015), it would be interesting to see if we could directly simulate the summary statistics which will reduce once again the computational burden.

Other contributions of the present dissertation focus on possible block approximations of the Gibbs distribution. These approaches allow to circumvent the major drawback of Monte Carlo approaches, namely their time complexity. As regards the estimation of posterior model parameter, we have illustrated the important role conditional composite likelihood approximations can play in the statistical analysis of Markov random field, and in particular Ising and autologistic models in spatial statistics, as a mean for overcoming the intractability of the likelihood function. In the Bayesian setting, we proposed an adjustment of the mode and the curvature at the mode of the posterior distribution as a way to correct the underestimated posterior mean and variance resulting from the use of a composite likelihood. This work has in particular pointed out that the main difficulty of such approaches is to handle the misspecification induced by non genuine likelihood functions.

In a context of model section, we proposed to move towards variational methods and in particular to use valid probability distributions in place of the intractable likelihood. Our proposal is to consider a product of valid distributions over non-overlapping

blocks of the lattice. Consequently, we derived Block Likelihood Information Criterion to discriminate between hidden Markov random fields. According to the numerical results of Chapter 4, the opportunity appears to be a satisfactory alternative to ABC model choice algorithms.

Perspectives

The first aspect of a further work is to scale-up all our approaches/algorithms to lattices of much larger size. Indeed, most of the results were carried out for regular 2D-grid of 100×100 sites. The flexibility and the ability of the approach to handle larger graph will play a decisive role in the design of useful methodology. The extension of the present work is not restricted to the only issue of larger graphs but can also be studied towards Gibbs random fields with larger number of parameters, such as the exponential random graph model particularly interesting for the analysis of social network. In the latter case, it could be interesting to adapt the methods of Chapter 2 to the hidden case and to study its performance as regards inference on random graph.

Another possibility yet to be explored is the robustness of the model choice approaches to the noise process. As a motivation of the question, we observed the limitation of the induced graph approach on the 16 colors example of Chapter 3 with a completely disordered noise. For continuous noise model, we have made two assumptions. The emission distribution was assumed to be first Gaussian and secondly conditionally independent. Whilst the latter modelling is quite standard, it would be interesting to further look noise process with heavier distribution tails such as the Student or Cauchy distributions as well as noise process that includes spatial information.

Besides modelling and practical limitations, the dissertation has raised some other questions. Our primary interest goes to the approximations of the Gibbs likelihood by probability distributions that factorize over blocks of the lattice. The latter approach offers an appealing trade-off between efficient computation and reliable results. Chapter 4 has introduced a novel criterion which makes in its current version two major approximations that are worth exploring. First mention, the choice of a particular substitute is lead by any optimality conditions. From that viewpoint, the construction of an optimal approximations regarding the variational free energy over the set of probability distributions that factorize on blocks is yet to be studied. At first sight, the existence of an explicit solution seems far from obvious. However it would be interesting to study their possible relation with *region-based approximations* of Yedidia et al. (2005). Their *generalized belief propagation* algorithms could give an estimate of the Kullback-Leibler divergence solution to build up the intended optimal approxi-

mations. Overall, variational methods, through the development of new calculation means, may offers reliable solution for hidden Markov random fields.

The second level of approximations concerns the penalty term. The next step of our work cannot be reduced to the sole aim of improving the quality of the approximations. Through Chapter 4, we have seen that an optimal solution with respect to the Kullback-Leibler divergence is not sufficient to ensure a good behaviour of model choice criteria, especially if the more complex model are not enough penalized. The penalty term used is solely valid for independent variable. We have neglected the interaction within a block, an assumption that slightly modified the number of free parameter. The impact of dependence variables on the penalty term is a logical follow-up to our work.

Bibliography

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. U. Asuncion, Q. Liu, A. T. Ihler, and P. Smyth. Learning with blocks: Composite likelihood and contrastive divergence. In *AISTATS, Journal of Machine Learning Research: W&CP*, volume 9, pages 33–40, 2010.

M. Baragatti and P. Pudlo. An overview on approximate Bayesian computation. *ESAIM: Proc.*, 44:291–299, 2014.

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.

M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 2009. doi: 10.1093/biomet/asp052.

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.

G. Biau, F. Cérou, and A. Guyader. New insights into Approximate Bayesian Computation. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques, in press*, 2013.

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.

A. Caimo and N. Friel. Bayesian model selection for exponential random graph models. *Social Networks*, 35(1):11 – 24, 2013.

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.

R. E. Chandler and S. Bate. Inference for clustered data using the independence loglikelihood. *Biometrika*, 94(1):167–183, 2007.

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.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

I. Csiszár and Z. Talata. Consistent Estimation of the Basic Neighborhood of Markov Random Fields. *The Annals of Statistics*, 34(1):123–145, 2006.

I. Csiszár, P. C. Shields, et al. The consistency of the BIC Markov order estimator. *The Annals of Statistics*, 28(6):1601–1619, 2000.

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.

P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

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.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

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.

P. Druilhet and J.-M. Marin. Invariant HPD credible sets and MAP estimators. *Bayesian Analysis*, 2(4):681–691, 2007.

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.

A. Estoup, E. Lombaert, J.-M. Marin, C. Robert, T. Guillemaud, P. Pudlo, and J.-M. Cornuet. Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. *Molecular Ecology Ressources*, 12(5):846–855, 2012.

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.

H. Geys, M. M. Regan, P. J. Catalano, and G. Molenberghs. Two Latent Variable Risk Assessment Approaches for Mixed Continuous and Discrete Outcomes from Developmental Toxicity Data. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(3):340–355, 2001.

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. A. Haughton. On the Choice of a Model to Fit Data from an Exponential Family. *The Annals of Statistics*, 16(1):342–355, 1988.

P. J. Heagerty and S. R. Lele. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111, 1998.

J. Heikkinen and H. Hogmander. Fully Bayesian approach to image restoration with an application in biogeography. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(4):569–582, 1994.

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.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine learning*, 37(2):183–233, 1999.

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

R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934, 1995.

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

J. T. Kent. Robust Properties of Likelihood Ratio Test. *Biometrika*, 69(1):19–27, 1982.

C. Keribin. Consistent Estimation of the Order of Mixture Models. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 62(1):49–66, 2000.

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.

B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1): 221–39, 1988.

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.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

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.

M. T. Moores, C. C. Drovandi, K. Mengersen, and C. Robert. Pre-processing for approximate Bayesian computation in image analysis. *Statistics and Computing*, 25(1): 23–33, 2015.

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.

D. J. Nott and T. Rydén. Pairwise likelihood methods for inference in image models. *Biometrika*, 86(3):661–676, 1999.

S. Okabayashi, L. Johnson, and C. J. Geyer. Extending pseudo-likelihood for Potts models. *Statistica Sinica*, 21(1):331, 2011.

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

F. Pauli, W. Racugno, and L. Ventura. Bayesian composite marginal likelihoods. *Statistica Sinica*, 21(1):149, 2011.

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.

P. Pudlo, J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, and C. P. Robert. ABC model choice via random forests. *ArXiv e-prints*, 1406.6288v2, 2014.

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.

A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.

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

M. Ribatet, D. Cooley, and A. Davison. Bayesian inference for composite likelihood models and an application to spatial extremes. *Statista Sinica*, 22:813–845, 2012.

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.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

T. Rydén. Consistent and asymptotically normal parameter estimates for hidden Markov models. *The Annals of Statistics*, 22(4):1884–1895, 1994.

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.

S. L. Scott. Bayesian Methods for Hidden Markov Models. *Journal of the American Statistical Association*, 97(457), 2002.

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.

L. Tierney and J. B. Kadane. Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

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.

S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

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.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282–2312, 2005.

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.

W. Zucchini and P. Guttorp. A Hidden Markov Model for Space-Time Precipitation. *Water Resources Research*, 27(8):1917–1923, 1991.

Résumé

La constante de normalisation des champs de Markov se présente sous la forme d'une intégrale hautement multidimensionnelle et ne peut être calculée par des méthodes analytiques ou numériques standards. Ceci constitue une difficulté majeure pour l'estimation des paramètres ou la sélection de modèle. Pour approcher la loi *a posteriori* des paramètres lorsque le champ de Markov est observé, nous remplaçons la vraisemblance par une vraisemblance composite, c'est à dire un produit de lois marginales ou conditionnelles du modèle, peu coûteuses à calculer. Nous proposons une correction de la vraisemblance composite basée sur une modification de la courbure au maximum afin de ne pas sous-estimer la variance de la loi *a posteriori*.
Ensuite, nous proposons de choisir entre différents modèles de champs de Markov cachés avec des méthodes bayésiennes approchées (ABC, *Approximate Bayesian Computation*), qui comparent les données observées à de nombreuses simulations de Monte-Carlo au travers de statistiques résumées. Pour pallier l'absence de statistiques exhaustives pour ce choix de modèle, des statistiques résumées basées sur les composantes connexes des graphes de dépendance des modèles en compétition sont étudiées à l'aide d'un taux d'erreur conditionnel original mesurant la puissance locale de ces statistiques à discriminer les modèles. Nous montrons alors que nous pouvons diminuer sensiblement le nombre de simulations requises tout en améliorant la qualité de décision, et utilisons cette erreur locale pour construire une procédure ABC qui adapte le vecteur de statistiques résumées aux données observées.
Enfin, pour contourner le calcul impossible de la vraisemblance dans le critère BIC (*Bayesian Information Criterion*) de choix de modèle, nous étendons les approches champs moyens en substituant la vraisemblance par des produits de distributions de vecteurs aléatoires, à savoir des blocs du champ. Le critère BLIC (*Block Likelihood Information Criterion*) que nous en déduisons permet de répondre à des questions de choix de modèle plus large que les méthodes ABC, en particulier le choix conjoint de la structure de dépendance et du nombre d'états latents. Nous étudions donc les performances de BLIC dans une optique de segmentation d'images.

Mots clefs: méthodes de Monte-Carlo, champs de Markov, statistique bayésienne, sélection de modèle, méthodes ABC, vraisemblances composites.

Abstract

Due to the Markovian dependence structure, the normalizing constant of Markov random fields cannot be computed with standard analytical or numerical methods. This forms a central issue in terms of parameter inference or model selection as the computation of the likelihood is an integral part of the procedure. When the Markov random field is directly observed, we propose to estimate the posterior distribution of model parameters by replacing the likelihood with a composite likelihood, that is a product of marginal or conditional distributions of the model easy to compute. Our first contribution is to correct the posterior distribution resulting from using a misspecified likelihood function by modifying the curvature at the mode in order to avoid overly precise posterior parameters.
In a second part we suggest to perform model selection between hidden Markov random fields with approximate Bayesian computation (ABC) algorithms that compare the observed data and many Monte-Carlo simulations through summary statistics. To make up for the absence of sufficient statistics with regard to this model choice, we introduce summary statistics based on the connected components of the dependency graph of each model in competition. We assess their efficiency using a novel conditional misclassification rate that evaluates their local power to discriminate between models. We set up an efficient procedure that reduces the computational cost while improving the quality of decision and using this local error rate we build up an ABC procedure that adapts the summary statistics to the observed data.
In a last part, in order to circumvent the computation of the intractable likelihood in the *Bayesian Information Criterion* (BIC), we extend the mean field approaches by replacing the likelihood with a product of distributions of random vectors, namely blocks of the lattice. On that basis, we derive BLIC (*Block Likelihood Information Criterion*) that answers model choice questions of a wider scope than ABC, such as the joint selection of the dependency structure and the number of latent states. We study the performances of BLIC in terms of image segmentation.

Key words: Monte-Carlo methods, Markov random fields, Bayesian statistics, model selection, approximate Bayesian computation, composite likelihood.