Calibration of conditional composite likelihood for Bayesian inference on Gibbs random fields

Julien Stoehr
Institut de Mathématiques
et de Modélisation de Montpellier,
UMR CNRS 5149,
Université de Montpellier, France.

Nial Friel
School of Mathematical Sciences and
Insight: the national center for data analytics,
University College Dublin, Ireland.

Abstract

Gibbs random fields play an important role in statistics, however, the resulting likelihood is typically unavailable due to an intractable normalizing constant. Composite likelihoods offer a principled means to construct useful approximations. This paper provides a mean to calibrate the posterior distribution resulting from using a composite likelihood and illustrate its performance in several examples.

1 Introduction

Gibbs random fields play an important and varied role in statistics. The autologistic model is used to model the spatial distribution of binary random variables defined on a lattice or grid (Besag, 1974). The exponential random graph model or p^* model is arguably the most popular statistical model in social network analysis (Robins et al., 2007). Other application areas include biology, ecology and physics.

Despite their popularity, Gibbs random fields present considerable difficulties from the point of view of parameter estimation, because the likelihood function is typically intractable for all but trivially small graphs. One of the earliest approaches to overcome this difficulty is the pseudolikelihood method (Besag, 1975), which replaces the joint likelihood function by the product of full-conditional distributions of all nodes. It is natural to consider generalizations which refine pseudolikelihood by considering products of larger collections of variables. The purpose of this paper is to consider such composite likelihood methods. In particular, we are interested in their use for Bayesian inference. Friel (2012) focused on a similar problem and studied how the size of the collections of variables influence the resulting approximate posterior distribution. Our main contribution is to present an approach to calibrate the posterior distribution resulting from using a mis-specified likelihood function.

This paper is organised as follows. Section 2 outlines a description of Gibbs random fields, and in particular the autologistic distribution. Composite likelihoods are introduced in Section 3. Here we focus especially on how to formulate conditional composite likelihoods for application to the autologistic model. We also focus on the issue of calibrating the composite likelihood function for use in a Bayesian context. Section 5 illustrates the performance of the various estimators for simulated data. The paper concludes with some remarks in Section 6.

2 Discrete-valued Markov random fields

A Markov random field y is a family of random variables y_i indexed by a finite set $\mathscr{S} = \{1, \ldots, n\}$ of nodes of a graph and taking values from a finite state space $\mathscr{Y}$. Here the dependence structure is given by an undirected graph $\mathscr{G}$ which defines an adjacency relationship between the nodes of $\mathscr{S}$: by definition i and j are adjacent if and only if they are directly connected by an edge in the graph $\mathscr{G}$. The likelihood of y given a vector of parameters $\theta = (\theta_1, \ldots, \theta_d)$ is defined as

$$f\left(y \mid \theta\right) \propto \exp(\theta^T s(y)) := q(y|\theta), \qquad (1)$$

where $s(y) = (s_1(y), \ldots, s_d(y))$ is a vector of sufficient statistics. However a major issue arises due to the fact that the normalizing constant in (1),

$$z(\theta) = \sum_{y \in \mathscr{Y}} \exp(\theta^T s(y)),$$

Appearing in Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. Copyright 2015 by the authors.

depends on the parameters θ, and is a summation over all possible realisation of the Gibbs random field. Clearly, $z(\theta)$ is intractable for all but trivially small situations. This poses serious difficulties in terms of estimating the parameter vector θ.

One of the earliest approaches to overcome the intractability of (1) is the pseudolikelihood method (Besag, 1975) which approximates the joint distribution of y as the product of full-conditional distributions for each y_i,

$$f_{\text{pseudo}}(y) = \prod_{i=1}^{n} f(y_i|y_{-i}, \theta),$$

where y_{-i} denotes $y \setminus \{y_i\}$. This approximation has been shown to lead to unreliable estimates of θ, see for example, Rydén and Titterington (1998), Friel et al. (2009). This is in fact one of the earliest composite likelihood approximations, and we will outline work in this area further in Section 3.

The autologistic model, first proposed by Besag (1972), is defined on a regular lattice of size $m \times m'$, where $n = mm'$. It is used to model the spatial distribution of binary variables, taking values $-1, 1$. The autologistic model is defined in terms of two sufficient statistics,

$$s_0(y) = \sum_{i=1}^{n} y_i, \quad s_1(y) = \sum_{j=1}^{n} \sum_{i \overset{\mathscr{G}}{\sim} j} y_i y_j,$$

where the notation $i \overset{\mathscr{G}}{\sim} j$ means that lattice point i is connected to lattice point j in $\mathscr{G}$. Following this notation, the normalizing constant of an autologistic model should be written $z(\theta, \mathscr{G})$, highlighting that it also depends on a graph of dependency. 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 y_i has neighbours $\{y_{i-m}, y_{i-1}, y_{i+1}, y_{i+m}\}$. Along the edges of the lattice each point has either 2 or 3 neighbours. The full-conditional of y_i can be written as

$$f(y_i|y_{-i}, \theta) \propto \exp(\theta_0 y_i + \theta_1 y_i (y_{i-m} + y_{i-1} + y_{i+1} + y_{i+m})), \quad (2)$$

where y_{-i} denotes y excluding y_i. As before, the conditional distribution is modified along the edges of the lattice. The Hammersley-Clifford theorem (Besag, 1974) shows the equivalence between the model defined in (2) and in (1). The parameter θ_0 controls the relative abundance of -1 and $+1$ values and the parameter θ_1 controls the level of spatial aggregation. Note that the Ising model is a special case, resulting from $\theta_0 = 0$.

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 $\{\mathscr{G}_1, \dots, \mathscr{G}_d\}$. Each graph of dependency $\mathscr{G}_k$ induces a summary statistic $s_k(y) = \sum_{j=1}^{n} \sum_{i \overset{\mathscr{G}_k}{\sim} j} y_i y_j$. For example, 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 y_i has neighbours $\{y_{i-1}, y_{i+1}\}$ according to $\mathscr{G}_1$ and $\{y_{i-m}, y_{i+m}\}$ according to $\mathscr{G}_2$. Along the edges of the lattice each point has either 1 or 2 neighbours. This allows to set an interaction strength that differs according to the direction.

3 Composite likelihoods

There has been considerable interests in composite likelihoods in the statistics literature. See, Varin et al. (2011) for a recent overview. Our primary objective is to work with a realisation from an autologistic distribution y. According to the previous section we denote $\mathscr{S} = \{1, \dots, mm'\}$ as an index set for the lattice points. Following Asuncion et al. (2010) we consider a general form of composite likelihood written as

$$f_{\text{CL}}(y \mid \theta) = \prod_{i=1}^{C} f(y_{A_i} \mid y_{B_i}, \theta).$$

Some special cases arise:

1. $A_i = A$, $B_i = \emptyset$, $C = 1$ corresponds to the full likelihood.
2. $B_i = \emptyset$ is often termed *marginal composite likelihood.*
3. $B_i = A \setminus A_i$ is often termed *conditional composite likelihood.*

The focus of this paper is on conditional composite likelihoods, since the autologistic distribution is defined in terms of conditional distributions. Note that the pseudolikelihood is a special case of 3. where each A_i is a singleton. 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 3., an exhaustive set of blocks would result in $C = (m - k + 1) \times (n - k + 1)$. In particular, we allow the collection of blocks $\{A_i\}$ to overlap with one another.

3.1 Bayesian inference using composite likelihoods

The focus of interest in Bayesian inference is the posterior distribution

$$p(\theta|y) \propto f\left(y \mid \theta\right) p(\theta). \tag{3}$$

Our proposal here is to replace the true likelihood $f\left(y \mid \theta\right)$ with a conditional composite likelihood, leading us to focus on the approximated posterior distribution

$$p_{\mathrm{CL}}(\theta \mid y) \propto f_{\mathrm{CL}}\left(y \mid \theta\right) p(\theta).$$

Surprisingly, there is very little literature on the use of composite likelihoods in the Bayesian setting, although Pauli et al. (2011) present a discussion on the use of conditional composite likelihoods. Indeed this paper suggests, following Lindsay (1988), that a composite likelihood should take the general form

$$f_{\mathrm{CL}}\left(y \mid \theta\right) = \prod_{i=1}^{C} f(y_{A_i} \mid y_{B_i}, \theta)^{w_i}, \tag{4}$$

where w_i are positive weights. In related work, Friel (2012) examined composite likelihood for various block sizes when $w_i = 1$. Our paper deals with the issue of calibrating the weights. Before focusing on the tuning of w_i, we highlight here the empirical observation that non-calibrated composite likelihood leads to an approximated posterior distribution with substantially lower variability than the true posterior distribution, leading to overly precise precision about posterior parameters, see Figure 1.

3.2 Computing full-conditional distributions of A_i

The conditional composite likelihood which we described above relies on evaluating

$$f(y_{A_i}|y_{-A_i}, \theta) = \frac{\exp\left(\theta_0 s_0(y_{A_i}) + s_1(y_{A_i} \mid y_{-A_i})\right)}{z(\theta, \mathscr{G}, y_{A_i})}, \tag{5}$$

where

$$s_0(y_{A_i}) = \sum_{j \in A_i} y_j, \quad s_1(y_{A_i}|y_{-A_i}) = \sum_{j \in A_i} \sum_{\ell \overset{\mathscr{G}}{\sim} j} y_\ell y_j.$$

Also the normalizing constant now includes the argument y_{A_i} emphasising that it involves a summation over all possible realisations of sub-lattices defined on the set A_i and conditioned on the realised y_{-A_i}, that is conditioned by all the lattice point of y_{-A_i} connected to a lattice point of y_{A_i} by an edge of $\mathscr{G}$. First we describe an approach to compute the overall normalizing constant for a lattice, without any conditioning on a boundary.

Generalised recursions for computing the normalizing constant of general factorisable models such as the autologistic models have been proposed by Reeves and Pettitt (2004). This method applies to autologistic lattices with a small number of rows, up to about 20, and is based on an algebraic simplification due to the reduction in dependence arising from the Markov property. It applies to un-normalized likelihoods that can be expressed as a product of factors, each of which is dependent on only a subset of the lattice sites. We can write $q(y \mid \theta)$ in factorisable form as

$$q(y \mid \theta) = \prod_{i=1}^{n} q_i(\boldsymbol{y}_i \mid \theta),$$

where each factor q_i depends on a subset $\boldsymbol{y}_i = y_i, y_{i+1}, \ldots, y_{i+m}$ of y, where m is defined to be the *lag* of the model. We may define each factor as

$$q_i(\boldsymbol{y}_i, \theta) = \exp\{\theta_0 y_i + \theta_1 y_i (y_{i+1} + y_{i+m})\} \tag{6}$$

for all i, except when i corresponds to a lattice point on the last row or last column, in which case y_{i+1} or y_{i+m}, respectively, drops out of (6).

As a result of this factorisation, the summation for the normalizing constant,

$$z(\theta, \mathscr{G}) = \sum_y \prod_{i=1}^{n} q_i(\boldsymbol{y}_i \mid \theta)$$

can be represented as

$$z(\theta, \mathscr{G}) = \sum_{y_n} q_n(\boldsymbol{y}_n \mid \theta) \cdots \sum_{y_1} q_1(\boldsymbol{y}_1 \mid \theta) \tag{7}$$

which can be computed much more efficiently than the straightforward summation over the 2^n possible lattice realisations. Full details of a recursive algorithm to compute the above can be found in Reeves and Pettitt (2004). Note that this algorithm was extended in Friel and Rue (2007) to also allow exact draws from $f(y|\theta)$

The minimum lag representation for an autologistic lattice with a first order neighbourhood occurs for r given by the smaller of the number of rows or columns in the lattice. Identifying the number of rows with the smaller dimension of the lattice, the computation time increases by a factor of two for each additional row, but linearly for additional columns. It is straightforward to extend this algorithm to allow one to compute the normalizing constant in (5), so that the summation is over the variables y_{A_i} and each factor involves conditioning on the set y_{-A_i}.

4 Bayesian composite likelihood adjustments

Approximating the true posterior distribution by remplacing the true likelihood by the composite likelihood leads to misspecification in the mean and variance of approximate posterior distribution as shown in Figure 1. The aim of the following Section is to establish identities that links the gradient and the Hessian of the log-posterior for θ to the moments of sufficient statistics with respect to the distribution of the Gibbs random field, whereupon we use these identities to calibrate the weights w_i in (4).

4.1 An estimation of the gradient and curvature of the posterior distribution

Using (3) as a starting point, we can write the gradient of the log-posterior for θ as

$$\nabla \log p\left(\theta \mid y\right) = s(y) - \nabla z(\theta, \mathscr{G}) + \nabla \log p(\theta).$$

It is straightforward to show that

$$\nabla z(\theta, \mathscr{G}) = \mathbb{E}_{y|\theta} s(y),$$

hence the gradient of the log-posterior for θ can be written as a sum of moments of $s(y)$, namely

$$\nabla \log p\left(\theta \mid y\right) = s(y) - \mathbb{E}_{y|\theta} s(y) + \nabla \log p(\theta). \qquad (8)$$

Taking the partial derivatives of the previous expression yields similar identity for the Hessian matrix of the log-posterior for θ,

$$\mathbf{H} \log p\left(\theta \mid y\right) = -\mathbf{K}_{y|\theta}(s(y)) + \mathbf{H} \log p(\theta), \qquad (9)$$

where $\mathbf{K}_{y|\theta}(s(y))$ denotes the covariance matrix of $s(y)$ when y has distribution $f\left(y \mid \theta\right)$. Similar to (8) and (9), one can express the gradient and Hessian of the log-posterior $\log p_{\mathrm{CL}}(\theta \mid y)$ in terms of moments of the sufficient statistics.

4.2 Mean adjustment

The mean adjustment aims to ensure that the posterior and the approximated posterior distributions have the same maximum. Thus, the adjustment here is simply the substitution

$$\overline{p_{\mathrm{CL}}}(\theta \mid y) = p_{\mathrm{CL}}(\theta - \theta^* + \theta^*_{\mathrm{CL}} \mid y),$$

where θ^* and θ^*_{CL}, is the maximum *a posteriori* (MAP) of the posterior distribution $p\left(\theta \mid y\right)$ and the approximated posterior distribution $p_{\mathrm{CL}}(\theta \mid y)$, respectively.

Addressing the issue of estimation of θ^* and θ^*_{CL}, we note generally from equation (9) that $\log p\left(\theta \mid y\right)$ and $\log p_{\mathrm{CL}}(\theta \mid y)$ are not concave functions. However the Hessian of the log-likelihood is a semi-negative matrix and so is unimodal. A reasonable choice of prior, for example with a semi-negative Hessian matrix, will thus lead to a unimodal posterior distribution. Care must be taken to ensure convergence of the optimisation algorithms to θ^* and θ^*_{CL}. In particular, we remark that since the approximate posterior distribution is typically very sharp around the MAP, as shown in Figure 1, it can be difficult to ensure convergence of gradient based algorithms in reasonable computational time. However, in our experiments we have found that using a BFGS algorithm which is based on a Hessian matrix approximation using rank-one updates calculated from approximate gradient evaluations, provided good performance in our context. Note that in practice, the gradient evaluated in the algorithm is stochastic and based on a standard Monte Carlo estimator of the expectation $\mathbb{E}_{y|\theta} s_j(y)$.

Algorithm 1: MAP estimation

```
Input: A lattice $y$
Output: Estimators $\widehat{\theta}^*$ of $\theta^*$ and $\widehat{\theta}^*_{\mathrm{CL}}$ of $\theta^*_{\mathrm{CL}}$

estimate $\theta^*_{\mathrm{CL}}$ using a BFGS algorithm based on
    Monte Carlo estimator of $\nabla \log p_{\mathrm{CL}}(\theta \mid y)$;
estimate $\theta^*$ using a BFGS algorithm based on
    Monte Carlo estimator of $\nabla \log p_{\mathrm{CL}}(\theta \mid y)$ and
    starting from $\widehat{\theta}^*_{\mathrm{CL}}$;
return $\widehat{\theta}^*$ and $\widehat{\theta}^*_{\mathrm{CL}}$;
```

Estimating $\widehat{\theta}^*$ using a random initialization point in BFGS algorithm (see Algorithm 1) is inefficient. Indeed, estimating $\mathbb{E}_{y|\theta} s(y)$ is the most cumbersome part of the algorithm and should be done as little as possible. Despite that $\widehat{\theta}^*_{\mathrm{CL}}$ is not equal to $\widehat{\theta}^*$ it is usually close and turns out to yield a good initialization to the second BFGS algorithm.

4.3 Magnitude adjustment

The general approach we propose to adjust the covariance of the approximated posterior is to temper the conditional composite likelihood with some weights w_i 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_i such that

$$\mathbf{H} \log p(\theta^* \mid y) = \mathbf{H} \log p_{\mathrm{CL}}(\theta^*_{\mathrm{CL}} \mid y).$$

Note in our context, there exists no particular reason to weight each blocks differently. Consequently we assume that each block has the same weight and we denote it w.

For the sake of simplicity, assume a uniform prior but everything can be easily written for any prior. When θ is a scalar parameter, writing identity (9) for $p\left(\theta \mid y\right)$

and $p_{\mathrm{CL}}(\theta \mid y)$ yields

$$w = \frac{\mathrm{Var}_{y\mid\theta^*}(s(y))}{\sum_{i=1}^{C} \mathrm{Var}_{y_{A_i}\mid\theta^*_{\mathrm{CL}}}(s(y_{A_i} \mid y_{-A_i}))}. \qquad (10)$$

However this approach does not apply when dealing with autologistic models since $\theta \in \mathbb{R}^d$ is a vector. We thus have a scalar constraint for an equality between the two matrices

$$\mathbf{K}_{y\mid\theta^*}(s(y)) = w \sum_{i=1}^{C} \mathbf{K}_{y\mid\theta^*_{\mathrm{CL}}}(s(y_{A_i} \mid y_{-A_i})).$$

In Table 1 we consider some possible identities that are natural to consider in order to choose a reasonable value for w. The options $w^{(3)}$ and $w^{(4)}$ include only the information contained in the diagonal of each matrix whereas options $w^{(1)}$, $w^{(2)}$ and $w^{(5)}$ take advantage of all the information of the covariance matrix.

Table 1: Weight options for a magnitude adjustment when $\theta \in \mathbb{R}^d$

$w^{(1)}$:	$\left\{\dfrac{\det\left[\mathbf{K}_{y\mid\theta^*}(s(y))\right]}{\det\left[\sum_{i=1}^{C} \mathbf{K}_{y\mid\theta^*_{\mathrm{CL}}}(s(y_{A_i} \mid y_{-A_i}))\right]}\right\}^{1/d}$
$w^{(2)}$:	$\dfrac{1}{d}\mathrm{tr}\left[\mathbf{K}_{y\mid\theta^*}(s(y))\left(\sum_{i=1}^{C} \mathbf{K}_{y\mid\theta^*_{\mathrm{CL}}}(s(y_{A_i} \mid y_{-A_i}))\right)^{-1}\right]$
$w^{(3)}$:	$\dfrac{1}{d} \cdot \sum_{i=1}^{d} \dfrac{\mathrm{Var}_{y\mid\theta^*}(s_i(y))}{\sum_{i=1}^{C} \mathrm{Var}_{y\mid\theta^*_{\mathrm{CL}}}(s_j(y_{A_i} \mid y_{-A_i}))}$
$w^{(4)}$:	$\dfrac{\mathrm{tr}\left[\mathbf{K}_{y\mid\theta^*}(s(y))\right]}{\mathrm{tr}\left[\sum_{i=1}^{C} \mathbf{K}_{y\mid\theta^*_{\mathrm{CL}}}(s(y_{A_i} \mid y_{-A_i}))\right]}$
$w^{(5)}$:	$\sqrt{\dfrac{\mathrm{tr}\left[\mathbf{K}^2_{y\mid\theta^*}(s(y))\right]}{\mathrm{tr}\left[\left(\sum_{i=1}^{C} \mathbf{K}_{y\mid\theta^*_{\mathrm{CL}}}(s(y_{A_i} \mid y_{-A_i}))\right)^2\right]}}$

4.4 Curvature adjustment

The adjustment presented in the previous Section only modify the magnitude of the approximated posterior but do not affect its geometry. The weight w similarly affects each direction of space parameters and does not take into account a possible modification of the correlation between the variables induced by the use of a composite likelihood approximation. 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 estimation of the abundance parameter and interaction parameter, θ_0 and θ_1, respectively, do not suffer from the same level of approximation relating to the independence assumption between blocks. Thus we should move from the general form (4) with a scalar weight on blocks to one involving a matrix of weights.

Following Ribatet et al. (2012) in the context of marginal composite likelihood, our strategy is to write

$$f\left(y \mid \theta\right) \approx f_{\mathrm{CL}}\left(y \mid \theta^*_{\mathrm{CL}} + W(\theta - \theta^*_{\mathrm{CL}})\right),$$

for some constant $d \times d$ matrix W. Note the substitution keeps the same maximum but deforms the geometry of the parameter space through the matrix W.

Assume that W is a lower triangular matrix in order to take into account the correlation between the parameter components. The suggestion of Ribatet et al. (2012) is to choose W in order to satisfy asymptotic properties of maximum composite likelihood estimators when the sample size tends to infinity. Since we only have one observation, we do not focus on the asymptotic covariance matrix results but rather on the covariance matrix at the estimated MAP. Indeed, we follow the same approach introduced in Section 4.3,

$$\mathbf{H} \log p(\theta^* \mid y) = \mathbf{H} \log p_{\mathrm{CL}}(\theta^*_{\mathrm{CL}} + W(\theta^* - \theta^*_{\mathrm{CL}}) \mid y),$$

which is equivalent to

$$\mathbf{H} \log p(\theta^* \mid y) = W^T \mathbf{H} \log p_{\mathrm{CL}}(\theta^* \mid y) W.$$

Note that the problem of uniqueness faced by Ribatet et al. (2012) due to a Cholesky decomposition does not exist here since we have access to a close form of different Hessians through Monte Carlo estimators. This leads to a system of equations that can be easily solved.

5 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 $z(\theta, \mathscr{G})$, $z(\theta, \mathscr{G}, y_{A_i})$ and to draw exactly from the distribution $f\left(y \mid \theta\right)$ or from the full-conditional distributions of A_i $f(y_{A_i} \mid y_{-A_i}, \theta)$. This exact computation of the posterior serves as a ground truth against which to compare with the posterior estimates of θ 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 algortihm 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 three 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.

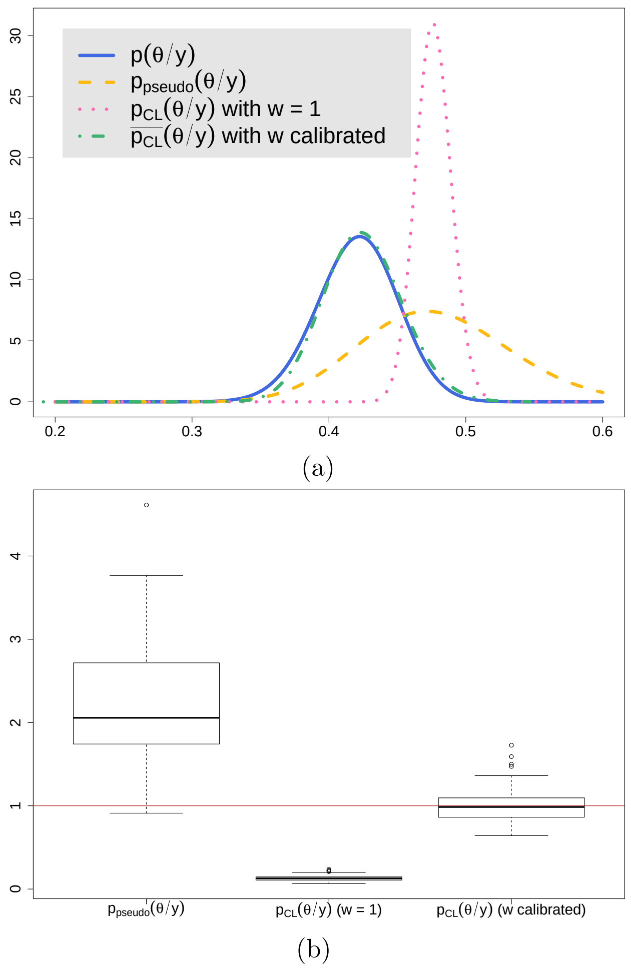
In each experiment, we simulated 100 realisations from the model. For each realisation, we use the BFGS algorithm 1 with an adhoc stopping condition to get the estimators $\widehat{\theta}^*$ and $\widehat{\theta}^*_{\mathrm{CL}}$. One iteration of the algorithm is based on a Monte Carlo estimator of either $\mathbb{E}_{y|\theta}s(y)$ or $\mathbb{E}_{y_{A_i}|\theta}s(y_{A_i} \mid y_{-A_i}, \theta)$ calculated from 100 exact draws whereas the Monte Carlo estimators of the covariance matrix $\mathbf{K}_{y|\widehat{\theta}^*}(s(y))$ and $\mathbf{K}_{y|\widehat{\theta}^*_{\mathrm{CL}}}(s(y_{A_i} \mid y_{-A_i}))$ are based on 50000 exact draws. In all experiments we placed uniform priors on θ.

Comparing the posterior $p(\theta \mid y)$ with the various posterior approximations $p_{\mathrm{CL}}(\theta \mid y)$ requires knowledge of the covariance matrix of θ. We could have used numerical integration but we prefered to use a simple MCMC algorithm. In terms of implementation, 7000 iterations were used with a burn in period of 2000 iterations for each dataset.

First experiment We considered the special case of a first-order Ising model with a single interaction parameter $\theta = 0.4$, 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 $\{\theta_k\}$ values, the right hand side of:

$$p(\theta_k \mid y) \propto \frac{q(y \mid \theta_k)}{z(\theta_k)} p(\theta_k),\ k = 1, \ldots, n,$$

can be evaluated exactly. Summing up the right hand side – using the trapezoidal rule – yields an estimate of the evidence, $p(y)$, which is the normalizing constant for the expression above and which in turn can be used to give a very precise estimate of $p(\theta \mid y)$. The plot so obtained for the posterior and posteriror approximations are given by Figure 1(a). On this example it should be clear that using an un-calibrated conditional composite likelihood leads to considerably underestimated posterior variances. But once we perform the mean adjusment and the magnitude adjusment, this provides a very good approximation of the true posterior. In Figure 1(b) we display the ratio $\mathbf{K}_{\mathrm{CL}}(\theta)/\mathbf{K}(\theta)$, where $\mathbf{K}(\theta)$, respectively $\mathbf{K}_{\mathrm{CL}}(\theta)$, denotes the variance of the posterior, respectively the posterior approximation, for θ, based on 100 realisations of a first-order Ising model. In view of these results there is no question that the magnitude adjustment (10) provides an efficient correction of the variance.



(a)

(b)

Figure 1: **First experiment results.** (a) Posterior distribution and posterior distribution approximations for θ of a first-order Ising model. (b) Boxplot displaying the ratio of the variances $\mathbf{K}_{\mathrm{CL}}(\theta)/\mathbf{K}(\theta)$ for 100 realisations of a first-order Ising model.

Table 2 confirms this result through evaluation of the relative mean square error $\mathbb{E}\left[(1 - \mathbf{K}_{\mathrm{CL}}(\theta)/\mathbf{K}(\theta))^2\right]$ and the average KL-divergence between the approximated posterior and true posterior distributions based on 100 realisations of a first order.

Second experiment We were interested in an anisotropic configuration of a first-order Ising model. We set $\theta = (0.3, 0.5)$. The evidence $p(y)$ is here estimated with an importance sampling method. We drew 1000 points using a Gaussian law whose moments are related to the Monte Carlo estimators of moments of

Table 2: Evaluation of the relative mean square error (RMSE) and the average KL-divergence (AKLD) between the approximated posterior and true posteriror distributions based on 100 simulations of a first-order Ising model.

COMP. LIKELIHOOD	RMSE	AKLD
$p_{\text{pseudo}}(\theta \mid y)$	1.96	0.510
$p_{\mathrm{CL}}(\theta \mid y)$ $(w = 1)$	0.757	0.337
$\overline{p_{\mathrm{CL}}}(\theta \mid y)$ (w defined by (10))	0.040	0.010

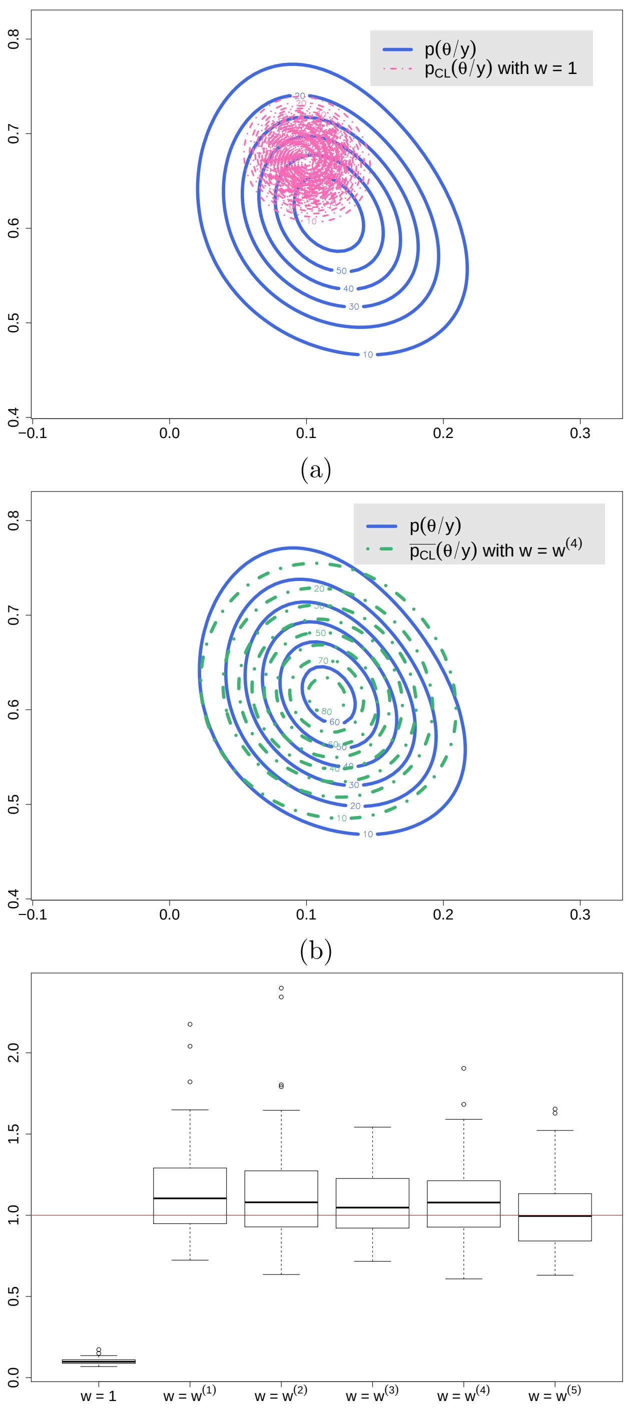
θ. Figure 2(a) and Figure 2(b) represent a comparison between the true likelihood and the estimates. As for the isotropic case, the mean and the magnitude adjustment allows us to build an accurate approximation of the posterior. In Figure 2(c) we display boxplots, based on 100 realisations of an anisotropic first-order Ising model, of the ratio $\|\mathbf{K}_{\mathrm{CL}}(\theta)\mathbf{K}^{-1}(\theta)\|_{\mathrm{F}}/\sqrt{2}$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. The different weight options are almost equivalent in term of variance correction. The weight w_5 seems to be the most informative. It should not be a surprise since it is based on the Frobenius norm which carries information of the matrix and its singular values.

This conclusion is emphasized in Table 3 which presents the relative mean square error $\mathbb{E}\left[\|1 - \mathbf{K}_{\mathrm{CL}}(\theta)\mathbf{K}^{-1}(\theta)\|_{\mathrm{F}}^{2}\right]$ and the average KL-divergence between the approximate and true posterior distributions for 100 realisations of the model.

Table 3: Evaluation of the relative mean square error (RMSE) the average KL-divergence (AKLD) between the approximated posterior and true posteriror distributions based on 100 simulations of an anisotropic first-order Ising model.

COMP. LIKELIHOOD	RMSE	AKLD
$p_{\mathrm{CL}}(\theta \mid y)$ $(w = 1)$	1.28	2.25
$\overline{p_{\mathrm{CL}}}(\theta \mid y)$ $(w = w^{(1)})$	0.555	0.067
$\overline{p_{\mathrm{CL}}}(\theta \mid y)$ $(w = w^{(2)})$	0.583	0.066
$\overline{p_{\mathrm{CL}}}(\theta \mid y)$ $(w = w^{(3)})$	0.540	0.071
$\overline{p_{\mathrm{CL}}}(\theta \mid y)$ $(w = w^{(4)})$	0.551	0.061
$\overline{p_{\mathrm{CL}}}(\theta \mid y)$ $(w = w^{(5)})$	0.525	0.079

Third experiment Here we focused on an autologistic model with a first-order dependance structure. The abundance parameter was set to $\theta_0 = 0.05$ and the interaction parameter to $\theta_1 = 0.4$. The differents implementation settings are exactly the same as for the second experiment. This example illustrates how the



(a)

(b)

(c)

Figure 2: **Second experiment results.**(a) Posterior distribution and posterior distribution approximation based on the conditional composite likelihood with $w = 1$. (b) Posterior distribution and posterior distribution approximation based on the conditional composite likelihood with mean and magnitude adjustments $(w = w^{(5)})$. (c) Boxplot displaying $\|\mathbf{K}_{\mathrm{CL}}(\theta)\mathbf{K}^{-1}(\theta)\|_{\mathrm{F}}/\sqrt{2}$ for 100 realisations of an anisotropic first-order Ising model.

use of composite likelihood approximation can induce a modification of the geometry of the distribution as shown in Figure 3(a). Indeed in addition to the mean and variance misspecification the conditional composite likelihood also changes the correlation between the variables. It should be evident that a magnitude adjustent would not be fruitful here since it would not affect the correlation. Instead the curvature adjustment manages to do so and thus yields a good approximation of the posterior, see Figure 3(b). One can object that we do not detect tail of the posterior. But Figure 3(c) and Table 4 show that the adjustment yields an efficient correction of the variance.

Table 4: Evaluation of the relative mean square error (RMSE) and the average KL-divergence (AKLD) between the approximated posterior and true posteriror distributions based on 100 simulations of a first-order autologistic model.

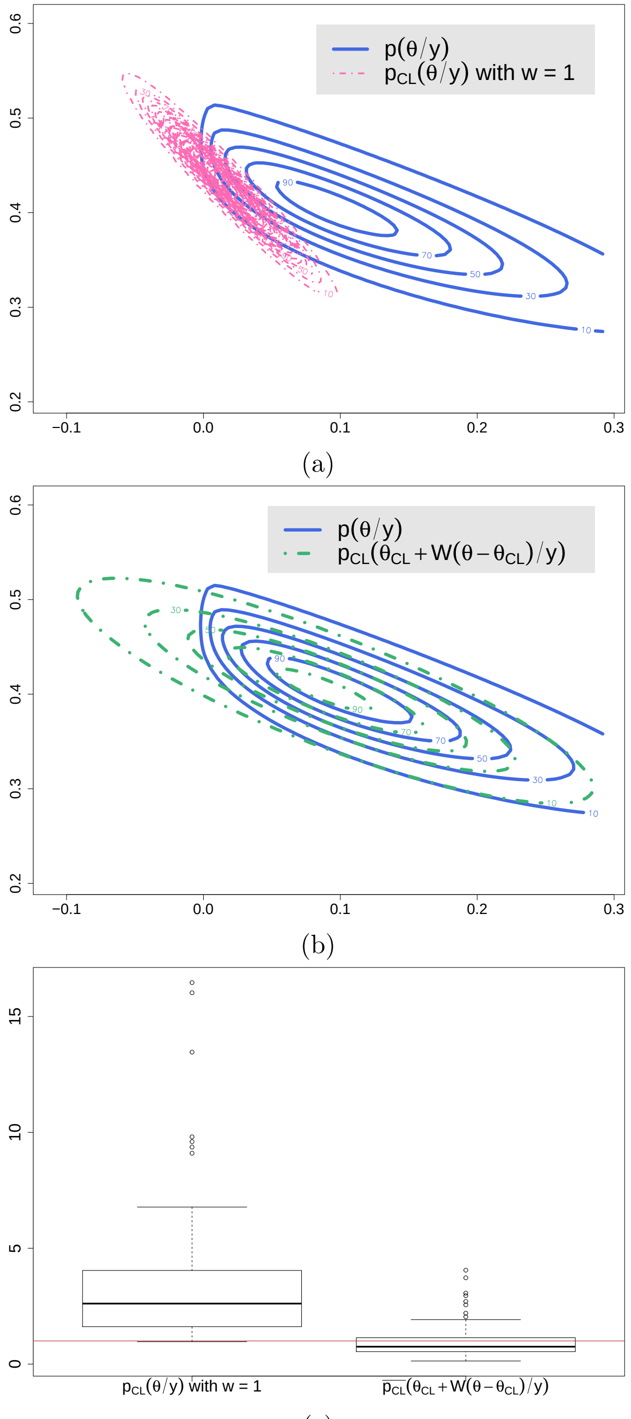
COMP. LIKELIHOOD	RMSE	AKLD
$p_{\mathrm{CL}}(\theta \mid y)\ (w=1)$	3.44	2.38
$p_{\mathrm{CL}}\left(\theta^*_{\mathrm{CL}} + W(\theta - \theta^*_{\mathrm{CL}}) \mid y\right)$	0.96	1.89

6 Conclusion

This paper has illustrated the important role that conditional composite likelihood approximations can play in the statistical analysis of Gibbs random fields, and in particular in the Ising and autologistic models in spatial statistics, as a means to overcoming the intractability of the likelihood function. However using composite likelihoods in a Bayesian setting can be problematic, since the resulting approximate posterior distribution is typically too concentrated and therefore underestimates the posterior mean and variance. Our main contribution has been to show how to calibrate the approximate posterior distribution that results from replacing the true likelihood with a conditional composite likelihood. Further work will focus on how to extend this framework to Gibbs random fields with larger number of parameters, such as the exponential random graph model.

Acknowledgments

We are grateful to Mathieu Ribatet and the anonymous reviewers of this paper for their helpful comments. The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289. Nial Friel's research was also supported by an Science Foundation Ireland grant: 12/IP/1424.



(a)

(b)

(c)

Figure 3: **Third experiment results.** (a) Posterior distribution and posterior distribution approximation based on the conditional composite likelihood with $w = 1$. (b) Posterior distribution and posterior distribution approximation based on the conditional composite likelihood with mean and curvature adjustments. (c) Boxplot displaying $\|\mathbf{K}_{\mathrm{CL}}(\theta)\mathbf{K}^{-1}(\theta)\|_{\mathrm{F}}/\sqrt{2}$ for 100 realisations of a first-order autologistic model.

References

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

J. 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. Nearest-neighbour systems and the autologistic model for binary data. *Journal of the Royal Statistical Society, Series B*, 34:75–83, 1972.

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

N. Friel. Bayesian inference for Gibbs random fields using composite likelihoods. In *Simulation Conference (WSC), Proceedings of the 2012 Winter*, pages 1–8, 2012.

N. Friel and H. Rue. Recursive computing and simulation-free inference for general factorizable models. *Biometrika*, 94:661–672, 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*, 18:243–261, 2009.

B. Lindsay. *Statistical inference from Stochastic processes*, volume 80, chapter Composite likelihoods, pages 221–239. American Mathematical Society, Providence, RI, 1988.

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.

F. Pauli, W. Racugno, and L. Ventura. Bayesian composite marginal likelihoods. *Statistica Sinica*, pages 149–164, 2011.

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

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph models for social networks. *Social Networks*, 29(2):169–348, 2007.

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

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

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