Likelihood Estimation and Inference for the Autologistic Model

N. FRIEL and A. N. PETTITT

The autologistic model is commonly used to model spatial binary data on the lattice. However, if the lattice size is too large, then exact calculation of its normalizing constant poses a major difficulty. Various different methods for estimation of model parameters, such as pseudo-likelihood, have been proposed to overcome this problem. This article presents a method to estimate the normalizing constant in an efficient manner. In particular, this allows tasks such as maximum likelihood estimation and inference for model parameters. We also consider the true likelihood approximated by the product of likelihoods for which the normalizing constant can be found by an analytic computational method by wrapping the lattice on the cylinder. This gives a simulation-free method of inference. We compare estimates of model parameters based on our new methods with the commonly used pseudo-likelihood approach. Although we have not considered Bayesian inferences here, the method can be straightforwardly extended to find posterior distributions. We apply our methods to the well-known endive data and to simulated data and find that our methods give substantially increased accuracy of estimation of model parameters.

Key Words: Autologistic distribution; Gibbs distribution; Ising model; Maximum likelihood; Markov chain Monte Carlo; Pseudo-likelihood.

1. INTRODUCTION

The autologistic model was first proposed by Besag (1972, 1974) and is widely used to model binary spatial data. A major drawback with its use, however, is that the normalizing constant is generally unknown analytically. In particular, this makes tasks such as maximum likelihood estimation impossible to carry out; see, for example, Huang and Ogata (2001). Many different methods have been introduced to overcome this problem by introducing approximate methods, for example, based on estimating equations, pseudo-likelihood or coding (Besag 1986). In fact pseudo-likelihood has been applied to a wide variety of spatial

---

N. Friel is Lecturer in Statistics, Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK (E-mail: nial@stats.gla.ac.uk). A. N. Pettitt is Professor, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia, (E-mail: a.pettitt@fsc.qut.edu.au).

Journal of Computational and Graphical Statistics, Volume 13, Number 1, Pages 232–246
DOI: 10.1198/1061860043029

data (see, e.g., Augustin, Mugglestone, and Buckland 1996; Wu and Huffer 1997). Geyer and Thompson (1992) provided a method for finding maximum likelihood estimates—Monte Carlo maximum likelihood—but this method, based on importance sampling ideas, is potentially very slow numerically and unstable due to calculation of the density and therefore exponentiation of large-valued statistics. The log-density is used in our methods, which are therefore numerically stable.

Pettitt and Friel (2001) showed how the normalizing constant for the autologistic model can be found by a computational analytic method if the lattice is wrapped onto the cylinder. But this method is restricted to lattices of size $m \times n$, $(m \leq n)$, with $m \leq 10$, as the method depends on finding eigenvalues of a $2^m \times 2^m$ matrix. Pettitt and Friel (2001) showed how it is possible to extend this result on the cylinder lattice to find the normalizing constant for the free boundary lattice using a Markov chain Monte Carlo scheme known as path sampling (Gelman and Meng 1998). This article shows that it is possible to extend these results to larger, arbitrary sized lattices. Further, we use this methodology to find the maximum likelihood estimate or posterior mode of model parameters for larger lattices and illustrate the method with the well known 14×179 lattice involving the endive data (Besag 1978) and simulated data.

For the endive data, we additionally find estimates and confidence intervals based on an approximation of the likelihood by approximating the normalizing constant as the product of two normalizing constants based on two 7×179 lattices found by splitting the original data into two sub-lattices. For each sub-lattice the normalizing constant is approximated by the corresponding normalizing constant for the cylinder. For the normalizing constants calculation we effectively treat the two sub-lattices as independent whereas the unnormalized distribution is taken to be its true value. We compare our two new methods of inference for the parameters with the well-known pseudo-likelihood approach. For the endive data we find good agreement for values of estimates and confidence intervals for our new methods, while the value of the pseudo-likelihood estimate is statistically significantly different from the maximum likelihood estimate. For simulated autologistic data there is a substantial difference between our new methods and the pseudo-likelihood approach. Our simulation results suggest that the true likelihood method and our good approximation to it are substantially more efficient than the pseudo-likelihood method. We find that the efficiency (defined as the ratio of the mean square error of the method and that of the true likelihood estimate) of the pseudo-likelihood estimate is typically 15% and our new approximate estimate about 60% efficient for the range of cases considered. These findings are also supported by Gu and Zhu (2001), who derived maximum likelihood estimates for the Ising model using an MCMC based Newton–Raphson algorithm. Additionally, our approximate methods can be used to find posterior distributions of parameters without recourse to simulation and only using MCMC to refine the normalizing constant approximation. However, here we illustrate the technique by finding confidence intervals for parameters based on the profile likelihood.

Section 2 briefly introduces the autologistic model and pseudo-likelihood estimation. Section 3 outlines the approach to calculate the normalizing constant for the lattice wrapped on the cylinder, and then shows how this result, when used with an MCMC scheme, may be

applied to calculate the normalizing constant for the free boundary lattice. Section 4 presents results of our maximum likelihood estimation procedures, together with confidence intervals based on profile likelihoods. Finally, we present conclusions to this work in Section 5.

2. AUTOLOGISTIC MODEL

Consider a binary random variable $\mathbf{x}_{ij}$ taking the values $\{0, 1\}$ at each site (i, j) on a regular $m \times n$ lattice. The unnormalized autologistic distribution on the lattice may be written in exponential form as

$$\begin{aligned} q(\mathbf{x}|\Theta) &= \exp\{\Theta^T V(\mathbf{x})\} \\ &= \exp\{\theta_0 V_0(\mathbf{x}) + \theta_f V_f(\mathbf{x})\}. \end{aligned} \tag{2.1}$$

Here $\Theta = (\theta_0, \theta_f)$ and $V(\mathbf{x}) = (V_0(\mathbf{x}), V_f(\mathbf{x}))$, with

$$V_0(\mathbf{x}) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \tag{2.2}$$

$$V_f(\mathbf{x}) = \sum_{i=1}^{m-1}\sum_{j=1}^{n} \mathbf{I}[x_{ij} = x_{i+1,j}] + \sum_{i=1}^{m}\sum_{j=1}^{n-1} \mathbf{I}[x_{ij} = x_{i,j+1}]. \tag{2.3}$$

Here $\mathbf{I}[x_{ij} = x_{i+1,j}]$ is the usual indicator function, taking the value 1 when $x_{ij} = x_{i+1,j}$, and 0 otherwise. It is seen that $V_0(\mathbf{x})$ is the number of 1's in the lattice. The first term of $V_f(\mathbf{x})$ counts the number of like-valued nearest neighbors within each column. The second term counts a similar number for within rows. Thus, $V_f(\mathbf{x})$ counts the number of like-valued direct adjacencies in the array. The subscript "f" denotes "free-boundary" lattice points. The statistics $V_0(\mathbf{x})$ and $V_f(\mathbf{x})$ are sufficient statistics for θ_0 and θ_f.

The normalizing constant is given by

$$z(\Theta) = z(\theta_0, \theta_f) = \int_{\mathbf{x}} q(\mathbf{x}|\Theta)\ \mu(d\mathbf{x}),$$

where μ is a counting measure and the integral becomes a sum over all possible outcomes of $\mathbf{x}$, in this case a sum involving 2^{mn} terms.

An important special case of the autologistic model occurs when the parameter $\theta_0 = 0$. This gives the well-known Ising model, studied extensively in statistical mechanics, and widely used as a prior distribution in Bayesian image analysis. In this instance, positive values of θ_f encourage patches of 0's or 1's and $\theta_f < 0$ promotes repulsive attraction between neighboring lattice points.

One of the major difficulties with the autologistic distribution is that its normalizing constant is often difficult to compute. Thus, in particular, maximum likelihood estimation of the model parameters, θ_0 and θ_f, becomes difficult. For this reason many different methods to estimate the likelihood have been proposed. The pseudo-likelihood method (Besag 1975) is one such example. We describe it in the subsequent section.

2.1 PSEUDO-LIKELIHOOD ESTIMATION

The pseudo-likelihood method proposed by Besag (1975) estimates the joint probability $P(\mathbf{x}|\Theta)$ as the product of the full conditionals. That is,

$$P(\mathbf{x}|\Theta) = \prod_{i=1}^{m}\prod_{j=1}^{n} P(x_{ij}|\mathbf{x}_{\setminus(i,j)}, \Theta). \tag{2.4}$$

Here $\mathbf{x}_{\setminus(i,j)}$ denotes the lattice $\mathbf{x}$ excluding point x_{ij}. Thus, pseudo-likelihood represents superficially only a slight departure from independence since the full conditionals are written as

$$P(x_{ij}|\mathbf{x}_{\setminus(i,j)}, \Theta) = \frac{\exp(\theta_0 x_{ij} + \theta_f \sum_{i,j\sim i',j'} \mathbf{I}[x_{ij} = x_{i'j'}])}{\exp(\theta_f \sum_{i,j\sim i',j'} \mathbf{I}[x_{ij} = 0]) + \exp(\theta_0 + \theta_f \sum_{i,j\sim i',j'} \mathbf{I}[x_{ij} = 1])}. \tag{2.5}$$

Here each summation is over first-order neighborhood structures, that is, summations over direct adjacencies. Again, $\mathbf{I}$ is the indicator function taking the value 1 when its argument is satisfied, and 0 otherwise.

It has been widely argued that pseudo-likelihood estimation gives misleading results, particularly when the interaction parameter θ_f is strong. For simulated data we find that pseudo-likelihood estimates are very inefficient compared with our full likelihood methods. We will return to this discussion later. Huang and Ogata (1999) use pseudo-likelihood estimates as the starting point of a single step Newton-Raphson Monte Carlo scheme to obtain better precision.

Finally, Huang and Ogata (2002) attempted to extend pseudo-likelihood estimation by considering products of blocks of lattices points conditioned on their compliments. As the blocks get smaller, the estimates should converge to pseudo-likelihood estimates. While as the blocks gets larger, the hope would be that the estimates get closer and closer to the true values of the maximum likelihood estimates.

3. LIKELIHOOD ESTIMATION FOR THE AUTOLOGISTIC MODEL

This section describes an approach to computationally efficient estimation of the normalizing constant of the autologistic distribution. Pettitt and Friel (2001, sec. 3) showed that if the autologistic distribution is wrapped on a cylinder, that is, when the first column and last column of the cylinder are direct neighbors of each other, then the corresponding normalizing constant can be exactly and efficiently computed. Let us outline the details here. The reader is referred to Pettitt and Friel (2001) for a more rigorous explanation.

3.1 NORMALIZING CONSTANT FOR THE CYLINDER LATTICE

We begin by extending the definition of the autologistic model as follows. Consider the unnormalized autologistic model defined as follows, with an extra parameter θ_c:

$$q(\mathbf{x}|\Theta) = \exp\{\theta_0 V_0(\mathbf{x}) + \theta_f V_f(\mathbf{x}) + \theta_c V_c(\mathbf{x})\}. \tag{3.1}$$

Here $V_0(\mathbf{x})$ and $V_f(\mathbf{x})$ are defined as before in (2.2) and (2.3), and now $\Theta = (\theta_0, \theta_f, \theta_c)$. The additional statistic $V_c(\mathbf{x})$ is defined as

$$V_c(\mathbf{x}) = \sum_{i=1}^{m} \mathbf{I}[x_{i,1} = x_{i,n}]. \tag{3.2}$$

So $V_c(\mathbf{x})$ is the product of direct adjacencies between the first and last columns, giving the so-called cylinder boundary conditions. Clearly, when $\theta_c = 0$ this model reduces to the standard autologistic model and when $\theta_c = \theta_f$ the model gives columns on the cylinder which have a stationary distribution.

The idea of the calculation of the normalizing constant for the cylindrical lattice, is to equate the n lattice columns, which are binary vectors of length m, to the outcomes over times $1, \ldots, n$ of a discrete vector stochastic process having its state space defined by all the possible arrangements of a binary m-vector, which number 2^m in total. The result basically derives from the realization that the normalizing constant is the trace of the product of n positive matrices, each similar to the transition matrix involving two adjacent times or columns. Further, as stated above, when the columns of the lattice are stationary then the product simplifies to be of the form $\mathbf{A}^T\mathbf{B}^n\mathbf{A}$ for $2^m \times 2^m$ matrices $\mathbf{A}$ and $\mathbf{B}$, and where $\mathbf{A}$ and the diagonal matrix $\mathbf{B}$ are derived from the one-step transition probability matrix for the binary m-vector. We therefore need to calculate the trace of $\mathbf{B}^n$, a diagonal matrix.

Note that this result is feasible for lattice sizes with smallest row or column less than or equal to 10. This is not necessarily a drawback, as we will see in the subsequent sections that we can use a Monte Carlo scheme to efficiently find normalizing constants for models which are related to the lattice with the cylinder boundary condition. For example, two $n \times m$ lattices can be "zipped" together to form a $n \times (2m)$ lattice, and so on. However, next we derive a simulation free approximation for the normalizing constant for larger lattices.

3.2 Simulation-Free Approximation of the Normalizing Constant

We now consider a new approximation based on the exact normalizing constant result for the cylinder boundary condition (Pettitt and Friel 2001). We first consider the standard identity $p(B_1, B_2) = p(B_1|B_2)p(B_2)$ with B_1, B_2 referring to the data for two sub-lattices B_1, B_2 of the original lattice $\mathbf{x}$. Given the Markov property of the lattice, the conditional distribution of B_1 given B_2 depends only on the sites of B_2 which constitute the boundary for B_1. If B_1 is relatively large in comparison to the boundary, then the conditional distribution $p(B_1|B_2)$ should be well approximated by the marginal $p(B_1)$. Or simply, we cut the lattice into sub-lattices which are treated as being independent. Then, as shown in Section 3.1, the normalizing constant can be found for lattices with smaller dimension being no more than 10, by using a computational analytic method. Hence likelihood inference can be carried out without simulation using the approximation that the normalizing constant for (B_1, B_2) is the product of the normalizing constants for B_1 and B_2. There is no need to approximate the unnormalized distribution $q(\mathbf{x}|\theta)$ as this is straightforwardly calculated. This approach

can obviously be extended to large lattices with many sub-lattices. We call this approach top×bottom estimation, and illustrate its performance in Section 4.

3.3 FROM CYLINDER TO FREE BOUNDARY LATTICE

This section shows how it is possible to extend the exact result for the normalizing constant on the cylindrical lattice to the free boundary lattice. To begin, note the following result which has appeared, for example, in Ripley (1988, p. 64) and Ogata (1989),

$$\log\left(\frac{z(\theta_0,\theta_f,\theta_a)}{z(\theta_0,\theta_f,\theta_b)}\right) = \int_{\theta_b}^{\theta_a} \mathbf{E}_{\mathbf{x}|(\theta_0,\theta_f,\theta_c)} V_c(\mathbf{x})\, d\theta_c. \tag{3.3}$$

This result shows that it is possible to explicitly calculate the ratio (or the log of the ratio) of two normalizing constants, when θ_c differs, and each of θ_0 and θ_f are fixed.

Our primary interest concerns calculating $z(\theta_0,\theta_f,\theta_c)$, when $\theta_c = 0$. However, we can use (3.3), and the exact result for the normalizing constant on the cylinder to do this since

$$\log\left(z(\theta_0,\theta_f,0)\right) = \log\left(\frac{z(\theta_0,\theta_f,0)}{z(\theta_0,\theta_f,\theta_f)}\right) + \log\left(z(\theta_0,\theta_f,\theta_f)\right).$$

Here the first term on the right-hand side can be calculated using (3.3), while the second expression on the right-hand side is the normalizing constant for the cylinder lattice, and so can be calculated exactly.

The expression (3.3) concerns an integral over $V_c(\mathbf{x})$, a sum of m terms, and provides a highly efficient way of evaluating $z(\theta_0,\theta_f,0)$. The reader is referred to Pettitt and Friel (2001, sec. 5) for a complete discussion, where the above conjecture is verified for various simulations.

Until now we have not mentioned the method of calculating the integral (3.3). Here we have some choices. One option—as described by Gelman and Meng (1998) and Pettitt and Friel (2001)—is the method of path sampling. Here the idea is to sample jointly from a distribution for $\mathbf{x}$ and θ with the two conditionals $p(\mathbf{x}|\theta) \propto q(\mathbf{x}|\theta)$ and $p(\theta|\mathbf{x}) \propto q(\mathbf{x}|\theta)$. This leads to the marginal $p(\theta) \propto z(\theta)$. If $z(\theta)$ does not change much in value over the range of θ, then this method is very efficient as the integral in (3.3) can be estimated by the simulation sum of generated values of $V_c(\mathbf{x})$ weighted inversely by the incremental differences in the ordered simulated values of θ (Gelman and Meng 1998, sec. 5). If $z(\theta)$ changes greatly over the range of θ, then the values of θ over this range are poorly sampled leading to a poor estimate of $\log z(\theta)$. In fact this is the situation we encountered in this study, and so we were prompted to look at alternative ways to estimate the integral (3.3).

A less efficient and more straightforward method is to estimate the integral using a quadrature rule. Simply choose a grid of θ_c values along the path of integration, and for each grid point, use MCMC theory to generate a Markov chain converging to $p(\mathbf{x}|\theta_0,\theta_f,\theta_c)$. Samples from this stationary distribution can then be used to estimate $\mathbf{E}_{\mathbf{x}|(\theta_0,\theta_f,\theta_c)} V_c(\mathbf{x})$ via an ergodic average. The integral (3.3) over θ_c can then be replaced by a discrete sum over the grid of θ_c values using a quadrature rule. We favor the trapezoidal rule.

3.4 EXTENSIONS TO LARGER LATTICE SIZES

As stated previously in Section 3.1, the exact result for the normalizing constant for the cylinder wrapped on a lattice may only be feasible for a lattice with smallest row or column less than or equal to 10. However, this section shows how this drawback may be overcome to estimate the normalizing constant for larger lattice sizes. We describe in detail the approach to calculate the normalizing constant for a lattice of size $(2m) \times n$, where $m < n$ and $m \leq 10$. It will become apparent how this may be generalized to larger lattice sizes.

Let us begin by reparameterizing the autologistic model assigning parameters to each of the $m \times n$ lattices in the top and bottom of the $(2m) \times n$ lattice. The sufficient statistic $V_0(\mathbf{x})$ can be written in terms of a statistic $V_{0,t}(\mathbf{x})$, defined for the upper half, "top", of the $(2m) \times n$ lattice with rows $i = 1, \ldots, m$, and $V_{0,b}(\mathbf{x})$, defined on the lower half, "lower", rows $i = m+1, \ldots, 2m$. So we have

$$V_0(\mathbf{x}) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} + \sum_{i=m+1}^{2m}\sum_{j=1}^{n} x_{ij} = V_{0,t}(\mathbf{x}) + V_{0,b}(\mathbf{x}).$$

In a like manner we can reparameterize the statistic $V_f(\mathbf{x})$. Here, however, there is a part, $V_z(\mathbf{x})$, common to both halves which involves elements in rows m and $m+1$:

$$\begin{aligned}
V_f(\mathbf{x}) &= \sum_{i=1}^{2m-1}\sum_{j=1}^{n} \mathbf{I}[x_{ij} = x_{i+1,j}] + \sum_{i=1}^{2m}\sum_{j=1}^{n-1} \mathbf{I}[x_{ij} = x_{i,j+1}] \\
&= \left(\sum_{i=1}^{m}\sum_{j=1}^{n-1} \mathbf{I}[x_{ij} = x_{i,j+1}] + \sum_{i=1}^{m-1}\sum_{j=1}^{n} \mathbf{I}[x_{ij} = x_{i+1,j}] \right) \\
&\quad + \sum_{j=1}^{n} \mathbf{I}[x_{mj} = x_{m+1,j}] \\
&\quad + \left(\sum_{i=m+1}^{2m}\sum_{j=1}^{n-1} \mathbf{I}[x_{ij} = x_{i,j+1}] + \sum_{i=m+1}^{2m-1}\sum_{j=1}^{n} \mathbf{I}[x_{ij} = x_{i+1,j}] \right) \\
&= V_{f,t}(\mathbf{x}) + V_z(\mathbf{x}) + V_{f,b}(\mathbf{x}), \qquad (3.4)
\end{aligned}$$

respectively.

Each of the newly defined parameters and corresponding statistics may be combined to give the unnormalized autologistic distribution

$$q(\mathbf{x}|\Theta) = \exp\{\theta_{0,b}V_{0,b}(\mathbf{x}) + \theta_{0,t}V_{0,t}(\mathbf{x}) + \theta_{f,b}V_{f,b}(\mathbf{x}) + \theta_z V_z(\mathbf{x}) + \theta_{f,t}V_{f,t}(\mathbf{x})\}$$

with

$$\Theta = (\theta_{0,b}, \theta_{0,t}, \theta_{f,b}, \theta_z, \theta_{f,t}).$$

Writing $z(\theta_{0,b}, \theta_{0,t}, \theta_{f,b}, \theta_z, \theta_{f,t})$ for the normalizing constant of this distribution, the following relationship holds for the normalizing constant of the $(2m) \times n$ lattice

$$z_{2m,n}(\theta_0, \theta_f) = z(\theta_0, \theta_0, \theta_f, \theta_f, \theta_f), \qquad (3.5)$$

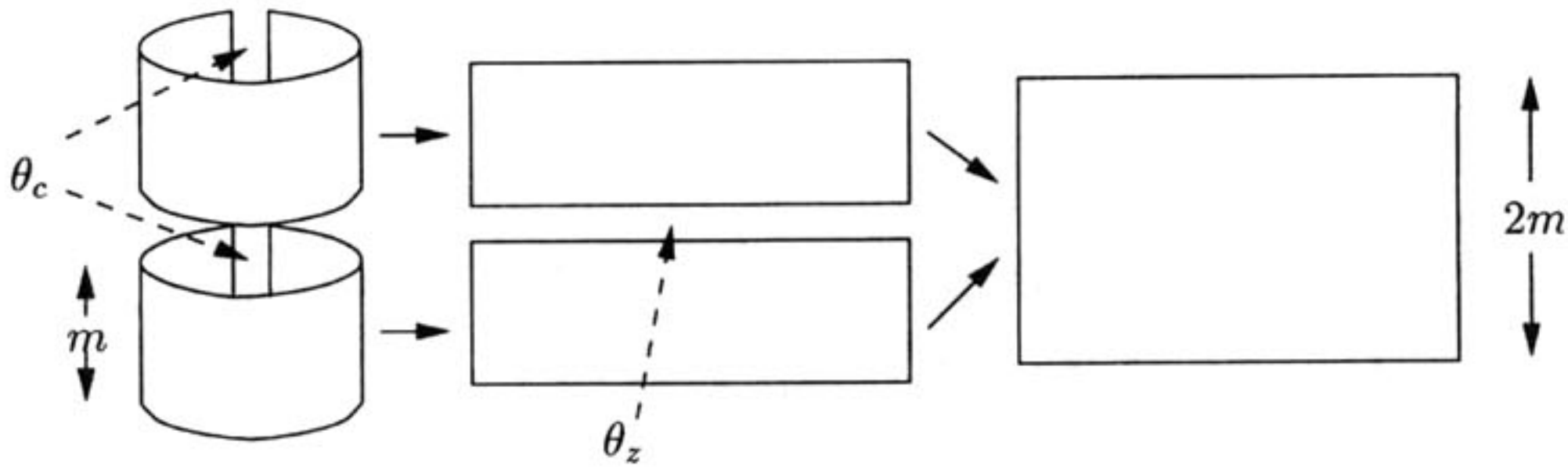

Figure 1. Evolution from two $m \times n$ lattices wrapped on a cylinder to a $(2m) \times n$ lattice.

and when $\theta_z = 0$ we can relate the normalizing constant of the $(2m) \times n$ lattice to the $m \times n$ lattice as follows

$$z(\theta_0, \theta_0, \theta_f, \theta_z = 0, \theta_f) = \{z_{mn}(\theta_0, \theta_f)\}^2 \tag{3.6}$$

because the top and bottom half of the lattice are made independent when θ_z is set equal to zero. Here we introduce subscripts to the notation for $z(\theta_0, \theta_f)$ to emphasize the corresponding lattice size, when this might be unclear. We can combine each of (3.5) and (3.6) in the following expression for the normalizing constant for the $(2m) \times n$ lattice:

$$\begin{aligned} z_{2m,n}(\theta_0, \theta_f) &= \frac{z(\theta_0, \theta_0, \theta_f, \theta_z = \theta_f, \theta_f)}{z(\theta_0, \theta_0, \theta_f, \theta_z = 0, \theta_f)} \times z(\theta_0, \theta_0, \theta_f, \theta_z = 0, \theta_f) \\ &= \frac{z(\theta_0, \theta_0, \theta_f, \theta_z = \theta_f, \theta_f)}{z(\theta_0, \theta_0, \theta_f, \theta_z = 0, \theta_f)} \times \{z_{mn}(\theta_0, \theta_f)\}^2. \end{aligned} \tag{3.7}$$

Consider the first expression on the right-hand side. Here only the θ_z parameter differs, when comparing the numerator and denominator, and so the log of its ratio may be estimated as an integral over θ_z, mimicking the discussion in Section 3.3. Because $z_{mn}(\theta_0, \theta_f)$ corresponds to a sufficiently small lattice it too may be easily calculated. Thus, taking logs we may rewrite (3.7) as

$$\log(z_{2m,n}(\theta_0, \theta_f)) = \int_0^{\theta_f} \mathbf{E}_{(\mathbf{x}|\theta_0, \theta_0, \theta_f, \theta_z, \theta_f)} V_z(\mathbf{x}) d\theta_z + 2\log(z_{mn}(\theta_0, \theta_f)). \tag{3.8}$$

To summarize, the procedure to calculate the $(2m) \times n$ lattice may be thought of as follows. Begin with two lattices of size $m \times n$ wrapped on a cylinder, and in each allow $\theta_c = 0$, to obtain two free boundary lattices of size $m \times n$. These two independent lattices are then joined together to form a $(2m) \times n$ lattice by introducing a parameter θ_z and corresponding statistic $V_z(\mathbf{x})$ which connects the smaller lattices along their last and first rows. Figure 1 aims to explain this idea graphically.

4. RESULTS

This section presents various methods to estimate likelihood parameters. We present what we term the *true likelihood* method which uses the exact estimate of the normalizing

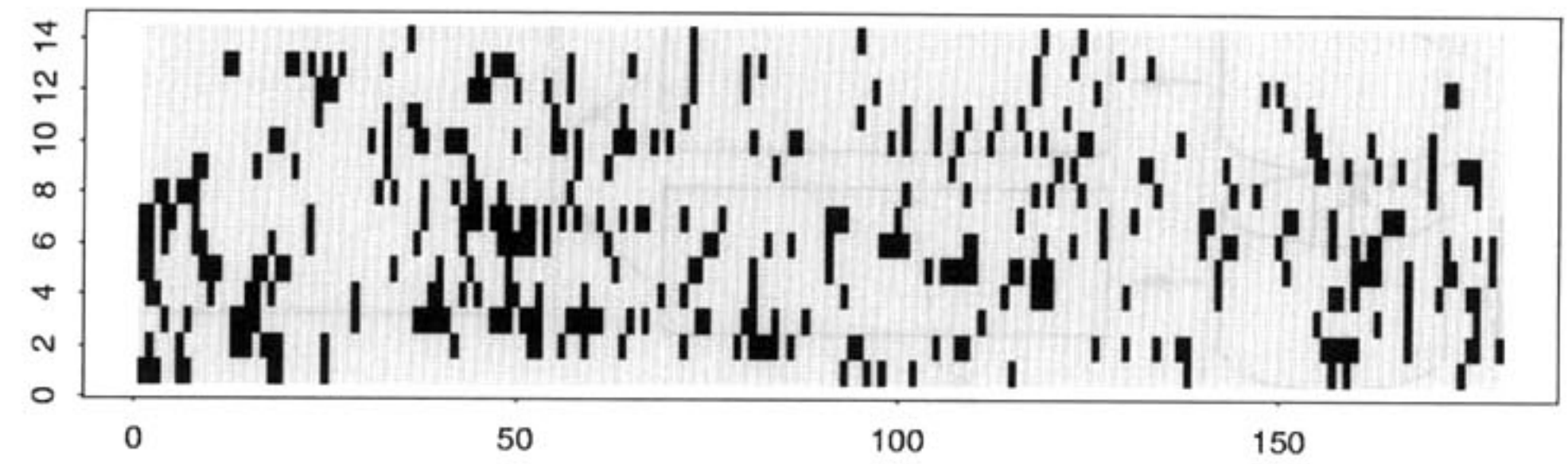

Figure 2. Endive dataset; see Besag (1978).

constant described above, together with the *top* $\times$ *bottom* estimation, described in Section 3.2. Essentially this amounts to splitting the lattice into two halves along its middle row. The full lattice is then treated as two independent lattices, each wrapped on a cylinder. Following the terminology of Section 3.4, this amounts to treating the variable θ_z as zero. Finally, both of these are compared with pseudo-likelihood estimates. Intuition would suggest that the top $\times$ bottom estimates should be closer to the true likelihood estimates than the pseudo-likelihood since the approximation involves an error for the log normalizing constant given by the integral in Equation (3.8) which, at its worst, should have a relative error of n in $2(m+1)n$ or 1 in $(2m+1)$.

The pseudo-likelihood and top $\times$ bottom estimates were found using simulated annealing. The normalizing constant for the true likelihood estimate was found as follows. First the interval $[0, \theta_f]$ in the integral (3.8) was discretized into 100 equal parts. An MCMC chain generated using Metropolis-Hastings updates with a burn-in of 5,000 iterations and 5,000 subsequent iteration values was used to estimate $\mathbf{E}(V_z(\mathbf{x}))$ at each point along the integral (3.8). The integral was then estimated by quadrature using the trapezoidal rule.

Clearly both pseudo-likelihood and top $\times$ bottom are simulation-free estimation methods, while the true likelihood method involves some simulation. It is a nontrivial matter to quantify bounds on the error of the simulation approximation. However, we note that perfect sampling (Propp and Wilson 1996) can be applied straightforwardly to the autologistic model, thus eliminating the error in estimating the burn in time to stationarity in each MCMC chain.

4.1 ENDIVE DATA

The following dataset first appeared in Besag (1978). It concerns the spread of footrot over an approximate regular lattice of size 14×179 of endive plants. These binary data describes the absence or presence of this disease. As before, we let $\mathbf{x} = \{x_{ij}\}$ describe the state of the lattice. Here $x_{ij} = 0$, if the plant is healthy, and 1, if the plant is diseased. Figure 2 displays this dataset. In this display black pixels correspond to lattice points with the value 1.

Table 1. Various Estimates for the Endive Dataset, Together with Corresponding Log-Likelihood Values.

	$\hat{\theta}_0$	$\hat{\theta}_f$	*Log-likelihood*
True likelihood	−0.801	0.389	−1053.96
Top×bottom	−0.803	0.401	−1055.82
Pseudo-likelihood	−0.781	0.398	−1061.71

We assume that this spatial dataset is a realization from an autologistic distribution. We return to this later. The main concern is inference for the model parameters. Table 1 presents three estimates: pseudo-likelihood estimates; estimates where the normalizing constant for the dataset is treated as two independent lattice each of size 7×179 (top×bottom); and finally the computationally efficient true likelihood estimates from Section 3. Each of these estimates is displayed with a log-likelihood value calculated using the exact method to calculate the normalizing constant.

The estimation procedure for the true likelihood method was repeated 10 times, using different initial random seeds. The corresponding estimates given in Table 1 are given as the average of these. The simulation standard errors for θ_0 and θ_f are 0.0007 and 0.0005, respectively.

Table 1 also shows that, indeed, the true likelihood estimate gives the largest log-likelihood value, not surprisingly, since this is how these estimates were chosen. Interestingly the estimates using the top × bottom method give a log-likelihood value closer to that of the true likelihood estimate, than for those estimated via pseudo-likelihood. This should agree with our intuition. Formally, if we test whether the pseudo-likelihood estimate values lie in a 95% confidence interval for the two parameters based on the true likelihood then, referring the deviance or twice the log-likelihood difference, $2(-1053.96 + 1061.71)$ or 15.5, to chi-squared on two degrees of freedom, we obtain a negative conclusion. Thus, the pseudo-likelihood estimate is certainly statistically different from the true maximum likelihood estimate in the sense that the value of the true likelihood evaluated at the pseudo-likelihood estimate is substantially less than the maximum of the likelihood.

We can find 95% confidence intervals based on the profile method using the likelihood or its approximation. For θ_0 and the true likelihood we obtain $[-0.851, -0.742]$, while for the top × bottom method we obtain $[-0.845, -0.746]$. For θ_f we obtain the intervals: $[0.365, 0.434]$ and $[0.377, 0.429]$, respectively, for the true and top×bottom methods. We note that the top×bottom confidence interval is somewhat more symmetric about the estimate than the true likelihood interval and this may be due to simualtion error in the estimation of the true likelihood normalizing constant.

4.2 SIMULATED DATA

This section presents results for simulated data. Here, for each choice of parameter values, we simulated 15 realizations of a lattice of size 12×100 from an autologistic distribution. To illustrate our findings we consider first the case with parameters $\theta_0 = 0$

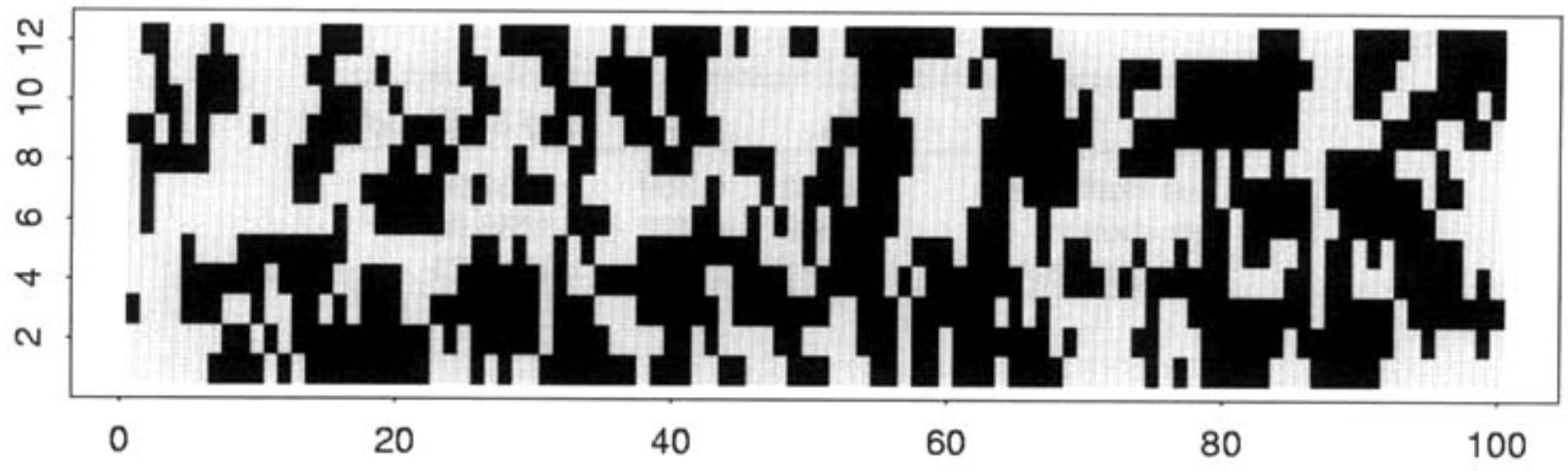

Figure 3. Simulated dataset, with $\theta_0 = 0$, $\theta_f = 0.3$.

and $\theta_f = 0.3$. For this choice, these are realisations from an Ising model. Figure 3 displays one such lattice. This plot shows characteristic behavior of the Ising model, namely, that there are distinct homogeneous regions of like valued points, since the interaction parameter $\theta_f = 0.3$ is quite large.

For each of the 15 lattices we have estimated maximum likelihood estimates of θ_0 and θ_f for the pseudo-likelihood, top×bottom and true likelihood methods. For each method we present the corresponding mean values together with corresponding standard errors. These may be seen in Table 2.

To compare the estimates we can consider the mean square errors of each method relative to that of the true likelihood method. Here the pseudo-likelihood estimation method has relative efficiencies in the range 5–17% compared with the true likelihood method, concurring with the view that the pseudo-likelihood can be unreliable and with the results of Gu and Zhu (2001) for the Ising model.

These results show that both the true likelihood and top×bottom methods perform better than the pseudo-likelihood method. The standard errors for both parameter values are considerably smaller than the corresponding standard errors for the pseudo-likelihood. The true likelihood method performs somewhat better than top×bottom, but at the expense of increased computational time. The top×bottom method has efficiencies in the range 52–60%. We also considered other values of the parameters and results are given in Table 3 and summary mean square error efficiencies given in Table 4. From Table 4 we note a consistent

Table 2. Means, Standard Errors, and Mean Squared Errors of Estimates of θ_0 and θ_f for 15 Realizations From an Autologistic Model With Parameters $\theta_0 = 0.0$ and $\theta_f = 0.3$, for Each of the Three Estimation Procedures

	True likelihood	*Top × bottom*	*Pseudo-likelihood*
Mean of $\hat{\theta}_0$	0.014	0.018	0.015
St. error for $\hat{\theta}_0$	0.00337	0.00441	0.01134
MSE for $\hat{\theta}_0$	0.000366	0.000616	0.002154
Mean of $\hat{\theta}_f$	0.304	0.297	0.310
St. error for $\hat{\theta}_f$	0.00216	0.00322	0.01007
MSE for $\hat{\theta}_f$	0.000086	0.000165	0.001621

Table 3. Means, Standard Errors, and Mean Squared Errors of Estimates of θ_0 and θ_f for 15 Realizations From an Autologistic Model with Various Different Parameter Values, for Each of the Three Estimation Procedures

		True likelihood	*Top × bottom*	*Pseudo-likelihood*
$\theta_0 = 0.0$	Mean of $\hat{\theta}_0$	−0.012	0.014	0.033
	St. error for $\hat{\theta}_0$	0.00293	0.00366	0.0104
	MSE for $\hat{\theta}_0$	0.000273	0.000397	0.00272
$\theta_f = -0.3$	Mean of $\hat{\theta}_f$	−0.291	−0.284	−0.285
	St. error for $\hat{\theta}_f$	0.00340	0.00399	0.0119
	MSE for $\hat{\theta}_f$	0.000254	0.000495	0.00234
$\theta_0 = 0.0$	Mean of $\hat{\theta}_0$	0.00100	0.0190	0.0230
	St. error for $\hat{\theta}_0$	0.00361	0.00354	0.00869
	MSE for $\hat{\theta}_0$	0.000196	0.000549	0.00166
$\theta_f = 0.1$	Mean of $\hat{\theta}_f$	0.0920	0.0940	0.0880
	St. error for $\hat{\theta}_f$	0.00416	0.00501	0.0128
	MSE for $\hat{\theta}_f$	0.000324	0.000413	0.00262
$\theta_0 = 0.0$	Mean of $\hat{\theta}_0$	−0.0200	−0.0180	−0.0210
	St. error for $\hat{\theta}_0$	0.00395	0.00417	0.0132
	MSE for $\hat{\theta}_0$	0.000634	0.000585	0.00304
$\theta_f = -0.1$	Mean of $\hat{\theta}_f$	−0.0930	−0.0910	−0.0810
	St. error for $\hat{\theta}_f$	0.00472	0.00518	0.0134
	MSE for $\hat{\theta}_f$	0.000383	0.000483	0.00489
$\theta_0 = 0.3$	Mean of $\hat{\theta}_0$	0.267	0.253	0.219
	St. error for $\hat{\theta}_0$	0.00869	0.00921	0.0163
	MSE for $\hat{\theta}_0$	0.00222	0.00348	0.0105
$\theta_f = 0.3$	Mean of $\hat{\theta}_f$	0.331	0.335	0.384
	St. error for $\hat{\theta}_f$	0.00929	0.0127	0.0173
	MSE for $\hat{\theta}_f$	0.00226	0.00664	0.0115
$\theta_0 = 0.3$	Mean of $\hat{\theta}_0$	0.326	0.343	0.393
	St. error for $\hat{\theta}_0$	0.00970	0.0117	0.0183
	MSE for $\hat{\theta}_0$	0.00209	0.00391	0.0137
$\theta_f = -0.3$	Mean of $\hat{\theta}_f$	−0.319	−0.322	−0.271
	St. error for $\hat{\theta}_f$	0.00881	0.00912	0.0173
	MSE for $\hat{\theta}_f$	0.00153	0.00173	0.00533

Table 4. Ratios of Mean Squared Errors for Both Top × Bottom and Pseudo-Likelihood to the Mean Squared Error for the True Likelihood

	$\frac{\text{MSE true likelihood}}{\text{MSE top}\times\text{bottom}}$	$\frac{\text{MSE true likelihood}}{\text{MSE pseudo-likelihood}}$
$\theta_0 = 0.0$	0.594	0.170
$\theta_1 = 0.3$	0.521	0.053
$\theta_0 = 0.0$	0.688	0.100
$\theta_1 = -0.3$	0.513	0.109
$\theta_0 = 0.0$	0.357	0.118
$\theta_1 = 0.1$	0.785	0.124
$\theta_0 = 0.0$	1.084	0.209
$\theta_1 = -0.1$	0.793	0.078
$\theta_0 = 0.3$	0.638	0.210
$\theta_1 = 0.3$	0.340	0.196
$\theta_0 = 0.3$	0.534	0.153
$\theta_1 = -0.3$	0.880	0.286

pattern across parameter values that the top×bottom method has efficiency around 60% with range 34–108% while pseudo-likelihood estimation has efficiency about 15% with range 5–29%. These results suggest that the top×bottom method is substantially superior to pseudo-likelihood without recourse to computer intensive simulation.

4.3 DISCUSSION

For the endive data, with quite extreme values of the parameters, estimates for the two model parameters differ by amounts which are statistically significantly different for the various methods. Comparing each of the estimates in terms of the log-likelihood values, calculated via the true likelihood method, it would appear that the top×bottom estimates are closer to the true likelihood estimates than those for the pseudo-likelihood estimates. While the top×bottom estimate is not statistically different from the true likelihood estimate (deviance difference 3.72 on two degrees of freedom) that of the pseudo-likelihood estimate is (deviance difference 15.5 on two degrees of freedom).

For simulated data we have found that the true likelihood and top×bottom methods both estimate model parameters very efficiently. Standard errors of these estimators, obtained from 15 simulated lattices with the same model parameters, were of the same order with the statistical efficiency (in terms of mean square error) of the top×bottom method in the range 34–108%. In comparison the pseudo-likelihood method estimated parameters with less precision, as illustrated by considerably higher standard errors of model parameters, having an efficiency in the range 5–29%.

It should be noted that for the endive data there may be some doubt as to the appropriateness of the autologistic model. For example, Besag (2000) suggested, using a Monte Carlo test, that the autologistic model gives a very poor fit to the endive data. This gives

more weight to the conclusions arising from the analysis of the simulated data, that the true likelihood method performs slightly better than the top×bottom method, at the expense of considerable computation time, and that both perform substantially better than the pseudo-likelihood method.

5. CONCLUSION

We have presented a computationally efficient but intensive method for calculating the normalizing constant for the autologistic model and incorporated this into a full likelihood analysis of binary spatial data allowing for estimation and inference for unknown model parameters. We have also presented a computationally efficient approximation to the true likelihood normalizing constant based on the product of two true likelihood normalizing constants, one for each split of the lattice into two sub-lattices. Finally, both of these methods are compared with the computationally fast estimation based on pseudo-likelihoods. The computational times for these methods vary in turns by orders of magnitude.

We suggest our remarks here would carry over to Bayesian analyses where use of the pseudo-likelihood could give misleading inference for model parameters. For example, Heikkinen and Hogmander (1994) used a single-parameter Ising model and pseudo-likelihood to carry out a so-called full Bayesian analysis of the presence/absence of Finnish toads which leads to substantial smoothing. Weir and Pettitt (2000) gave an alternative full Bayesian analysis of the data which leads to little spatial smoothing.

In short we suggest that our methods provide improvements to the widely used pseudo-likelihood approach giving substantially more accurate inference, but at the price of some computational resources.

ACKNOWLEDGMENTS

The authors thank Julian Besag for providing the endive data. Nial Friel was supported by an Australian Research Council grant. We would like to thank the referees for constructive comments which have considerably improved this article.

[Received December 2000. Revised August 2002.]

REFERENCES

Augustin, N., Mugglestone, M., and Buckland, S. (1996), "An Autologistic Model for Spatial Distribution of Wildlife," *Journal of Applied Ecology*, 33, 339–347.

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

——— (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society*, Series B, 36, 192–236

——— (1975), "Statistical Analysis of Non-lattice Data," *The Statistician*, 24, 179–195.

——— (1978), "Some Methods of Statistical Analysis for Spatial Data," *Bulletin of the International Statistical Institute*, 47, 77–92.

——— (1986), "On the Statistical Analysis of Dirty Pictures" (with discussion), *Journal of the Royal Statistical Society*, Series B, 48, 259–302.

——— (2000), "An Introduction to Markov Chain Monte Carlo," unpublished report.

Gelman, A., and Meng, X.-L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

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

Gu, M. G., and Zhu, H.-T. (2001), "Maximum Likelihood Estimation for Spatial Models by Markov Chain Monte Carlo Stochastic Approximation," *Journal of the Royal Statistical Society*, Series B, 63, 339–355.

Heikkinen, J., and Hogmander, H. (1994), "Fully Bayesian Approach to Image Restoration With an Application in Biogeography," *Applied Statistics*, 43, 569–582.

Huang, F., and Ogata, Y. (1999), "Improvements of the Maximum Pseudo-Likelihood Estimators in Various Spatial Statistical Models," *Journal of Computational and Graphical Statistics*, 8, 510–530.

——— (2001), "Comparison of Two Methods for Calculating the Partition Function of Various Spatial Statistical Models," *Australian and New Zealand Journal of Statistics*, 43, 47–65.

——— (2002), "Generalized Pseudo-Likelihood Estimates for Markov Random Fields on a Lattice," *Annals of the Institute of Statistical Mathematics* (to appear).

Ogata, Y. (1989), "A Monte Carlo Method for High-Dimensional Integration," *Numerical Mathematics*, 55, 137–157.

Pettitt, A. N., Friel, N., and Reeves, R. (2003), "Efficient Calculation of the Normalisation Constant of the Autologistic Model on the Lattice," *Journal of the Royal Statistical Society*, Series B, 65, 235–247.

Propp, J. G., and Wilson, D. B. (1996), "Exactly Sampling with Coupled Markov Chains and Applications to Statistical Mechanics," *Random Structures and Algorithms*, 9, 223–252.

Ripley, B. D. (1988), *Statistical Inference for Spatial Processes*, New York: Cambridge University Press.

Weir, I. S., and Pettitt, A. N. (2000), "Binary Probability Maps Using a Hidden Conditional Autoregressive Gaussian Process with an Application to Finnish Common Toad Data," *Applied Statistics*, 49, 473–484.

Wu, H., and Huffer, F. W. (1997), "Modelling the Distribution of Plant Species Using the Autologistic Regression Model," *Ecological Statistics*, 4, 49–64.