

An Adjustment for Edge Effects Using An Augmented Neighborhood Model in the Spatial Auto-logistic Model

Johan Lim, Xinlei Wang, and Michael Sherman *

Abstract

Parameter estimation in the spatial auto-regressive models has difficulty from the edge sites which have unobserved neighborhood sites. Some ad hoc remedies suggested in the literature are the free boundary condition, the toroidal boundary condition, or estimation with only the data for internal sites. However, parameter estimates are often sensitive to the assumption for the unobserved neighborhood sites and all the above assumptions have some apparent shortcomings such as systematic bias or inflated variance. In this paper, we propose a new way to incorporate the edge sites by introducing an augmented random neighborhood, denoted by the augmented neighborhood model, which represents the entire external field. To estimate the model, we derive the EM procedures for the maximum pseudo likelihood estimator and the maximum likelihood estimator. Several simulation studies show that the random external field provides better performance of the maximum pseudo-likelihood estimator and the maximum likelihood estimator than other assumptions to the edge sites. As an example, we apply the random external field to modeling the distribution of *Plantago lanceolata* in Kansas.

*Johan Lim is at Department Applied Statistics, Yonsei Univeristy, and on leave from Department of Statistics, Texas A&M University; Michael Sherman is from Department of Statistics, Texas A&M University; and Xinlei Wang is from Department of Statistical Science, Southern Methodist University. All correspondences are to Johan Lim, Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843-3143, johanlim@stat.tamu.edu

1 Introduction

Conditional autoregressive models (CAR) play a crucial role for spatial systems in many disciplines such as engineering (Geman and Geman, 1984), social science (Anselin and Cho, 2003) and ecology (Legendre and Fortin, 1989), where spatially correlated observations are encountered. Parameter estimation in the CAR is often a fundamental and complex step towards statistical inferences and has been addressed by various authors (*e.g.*, Besag 1975, Pickard 1982 and 1987, Younes 1991, Comets 1992, Geyer and Thompson 1992, Guyon and Künsch 1992, Huang and Ogata 1999 and 2002, Bartolucci and Besag 2002, Reeves and Pettitt 2004). In this paper, we investigate an important issue closely related to the problem of parameter estimation, namely, the edge effects, *i.e.*, spatial locations that do not have fully observed neighbors.

Consider a fixed system of nodes, labeled by a set D of location, and an associated vector Z of (dependent) random variables Z_i with binary responses $-1/1$ for $i \in D$. For each site i , let $N(i)$ denote the set of its “neighbors” and Z_{-i} denote the vector of Z_j ’s where $j \in D$ and $j \neq i$. Let ∂D be the set of indices for the boundary some of whose neighbors are not in D . The conditional auto–logistic model (the most common CAR model for binary observations) is locally defined as:

$$\mathbb{P}\left(Z_i = z_i \mid Z_{-i} = z_{-i}\right) = \frac{\exp\left\{z_i\left(\alpha + \beta \sum_{j \in N(i)} z_j\right)\right\}}{\exp\left\{\alpha + \beta \sum_{j \in N(i)} z_j\right\} + \exp\left\{-\alpha - \beta \sum_{j \in N(i)} z_j\right\}}, \quad (1)$$

where the parameter α determines the overall proportion of Z_i ’s with value one and the parameter β reflects the intensity of the interaction between Z_i and its neighborhood (Besag, 1974). Note that for $i \in \partial D$, $N(i)$ contains unobserved components which are not in D so that (1) is not well defined for parameter estimation. This difficulty occurs not only for the auto–logistic model, but also for all CAR models including the autoregressive Poisson or Gaussian models. In the sequel, we focus on the binary case but comment on

the Gaussian case in the discussion.

To resolve the difficulty from the boundary sites, most previous work is based on some ad-hoc treatment of the unobserved external locations, which are often computationally convenient but have no solid empirical or theoretical justification. Such approaches include the free boundary condition (FREE), the toroidal boundary condition (TORUS), and omitting all boundary observations (INNER). The free boundary condition assumes that all nodes in the external field have a value of zero and the toroidal boundary condition assumes the spatial system as a “torus”. A more rigorous approach to adjust for the boundary sites is to integrate out the unobserved nodes with respect to its conditional distribution given the observed values (*e.g.*, p.439 in Cressie, 1993). However, it presumes the knowledge on the system of the external field which is often not available in practice.

In this paper, we propose a data-driven approach that treats the entire external field as *one unobserved random variable*, and then estimates it jointly with the parameters of the model. The proposed augmented neighborhood variable approach can be considered on the same line with the elaborative approach discussed in Cressie (1993) but it has at least two of its own merits. First, the proposed augmented variable approach does not assume the structure of the external field and can be applied to both lattice and non-lattice system. Second, in comparison with the TORUS, the augmented neighborhood model introduces one random component in bridging one end and the other end. We expect that this random layer interweaves two ends which may be quite different. To estimate the models, we derive the EM procedures (Dempster, Laird and Rubin 1977) for the MPLE and the MLE along with the unobserved augmented neighborhood variable. We implement several simulation studies which demonstrate that the maximum likelihood estimator (MLE) and the maximum pseudo likelihood estimator (MPLE) with the augmented neighborhood perform better than those with other assumptions on the external

field. In addition, we apply the to modeling the distribution of the *Plantago lanceolata* distribution in Kansas and compare the results with those from other existing approaches. Straightforward extensions to other CARs will be discussed in the Conclusion.

The paper is organized as follows. In Section 2, we shortly review two basic estimation procedures for the CAR, the MLE and the MPLE, and several existing treatments to the edge sites. Section 3 introduces the augmented neighborhood model and derives the EM procedures to evaluate the MLE and the MPLE. Section 4 implements simulation studies to evaluate the performance of the MLE and the MPLE with the ANV and to compare it with other alternative treatments. In Section 5, we apply the procedures with the augmented neighborhood model to modeling the *Plantago lanceolata* distribution in Kansas. Finally, Section 6 concludes the paper with the discussion of the extension to other CARs.

2 Two Existing Estimation Procedures and Edge Effects

In recent years, the problem of parameter estimation for spatial auto-logistic models has attracted considerable statistical attention. For a detailed review and discussion, see, *e.g.*, Sherman et al. (2005) and the references therein. In this section, we briefly describe two basic methods ML and MPL where most existing estimation procedures are grounded. Then we focus on several common treatments to the boundary problem associated with each method. As will be shown next, the parameter estimation in the spatial auto-logistic models is much influenced by the assumption to the external field, especially when the size of the system is not sufficiently large.

2.1 Maximum Likelihood Estimator

As shown in Cressie (1993), the conditional auto–logistic model defined in (1) has the joint probability distribution function

$$\mathbb{P}(Z) = \frac{1}{\Psi(\alpha, \beta)} \exp \left\{ \alpha \sum_{i \in D} Z_i + \frac{\beta}{2} \sum_{i \in D} Z_i \left(\sum_{j \in N(i)} Z_j \right) \right\}, \quad (2)$$

where $\Psi(\alpha, \beta)$ is the partition function defined as

$$\Psi(\alpha, \beta) = \sum_{Z} \exp \left\{ \alpha \sum_{k \in D} Z_k + \frac{\beta}{2} \sum_{k \in D} Z_k \left(\sum_{\ell \in N(k)} Z_\ell \right) \right\}. \quad (3)$$

Note the summation in (3) is over all possible realizations of Z .

Equivalently, we can denote $\mathbb{P}(Z)$ in (2) as $L(\alpha, \beta; Z)$, the likelihood function of the data. Maximization of $L(\alpha, \beta; Z)$ is, in general, computationally impractical when the total number of sites in D is moderate or large because of the large number of terms in $\Psi(\alpha, \beta)$ that must be summed. In particular, \sum_Z is over $2^{|D|}$ where $|D|$ denotes the number of sites in D . To avoid this difficulty in calculating the MLE, many authors have proposed to approximate $L(\alpha, \beta; Z)$ or its derivatives via Monte Carlo samples (*e.g.*, Besag, 1986; Geman and Geman, 1984; Younes, 1991; Moyeed and Baddeley, 1991; Geyer and Thompson, 1992; Huang and Ogata, 1998). However, people confront the boundary problem when generating the Monte Carlo samples to approximate the MLE. To overcome this, there are two choices:

- Assume free boundary condition that all nodes in the external field have zero values. It is denoted as FREE in this paper.
- Assume a "periodic boundary" and make the observed spatial system as a self-closed system (*i.e.*, the neighbors of each pixel are in the system); Or equivalently, construct a torus. It is denoted as TORUS in this paper.

The former, with the free boundary condition, can be applied to both lattice and non-lattice system whereas the latter is only available to a lattice spatial system, which is very restrictive for many applications. Also, the torodial boundary condition tends to systematically underestimate the interaction intensity $|\beta|$. This can be seen by observing that, a torus system assumes that $Z_{0,j} = Z_{m,j}$, $Z_{m+1,j} = Z_{1,j}$, $Z_{i,0} = Z_{i,n}$, and $Z_{i,n+1} = Z_{i,1}$ for all i and j in an $m \times n$ lattice; using sites on one side as the neighbors of those on the other side artificially lessens the interaction on the boundary because more distant sites are typically less interacted.

2.2 Maximum Pseudo Likelihood Estimator

Besag (1974) notes that $\mathbb{P}(Z_i = z_i | Z_{-i})$ and $\mathbb{P}(Z_j = z_j | Z_{-j})$ are conditionally independent if $i \notin N(j)$ and $j \notin N(i)$. Accordingly, the data grid can be “coded” into two groups of observations such that within each group individual components are conditionally independent. Thus, the usual likelihood theory for the data in each group applies. Ignoring the dependence between the two groups and pooling all components together form the so-called “pseudo likelihood function” (Besag, 1975):

$$\text{PL}(\alpha, \beta; Z) = \prod_{i \in D} \frac{\exp \{Z_i(\alpha + \beta \sum_{j \in N(i)} Z_j)\}}{\exp \{\alpha + \beta \sum_{j \in N(i)} Z_j\} + \exp \{-\alpha - \beta \sum_{j \in N(i)} Z_j\}}. \quad (4)$$

As mentioned in Possolo (1986) and Sherman et al. (2005), the MPL estimate can be computed from a standard logistic regression with response variables $\mathbf{y} = (Z_i + 1)/2$ and corresponding covariate $\mathbf{x} = \sum_{j \in N(i)} Z_j$. Further, the consistency and the asymptotic normality of the MPLE have been demonstrated by several authors (*e.g.*, Geman and Graffigne, 1986; Comets, 1992; Guyon and Künsch, 1992), and the concept of the PL has been applied to many other problems without difficulty.

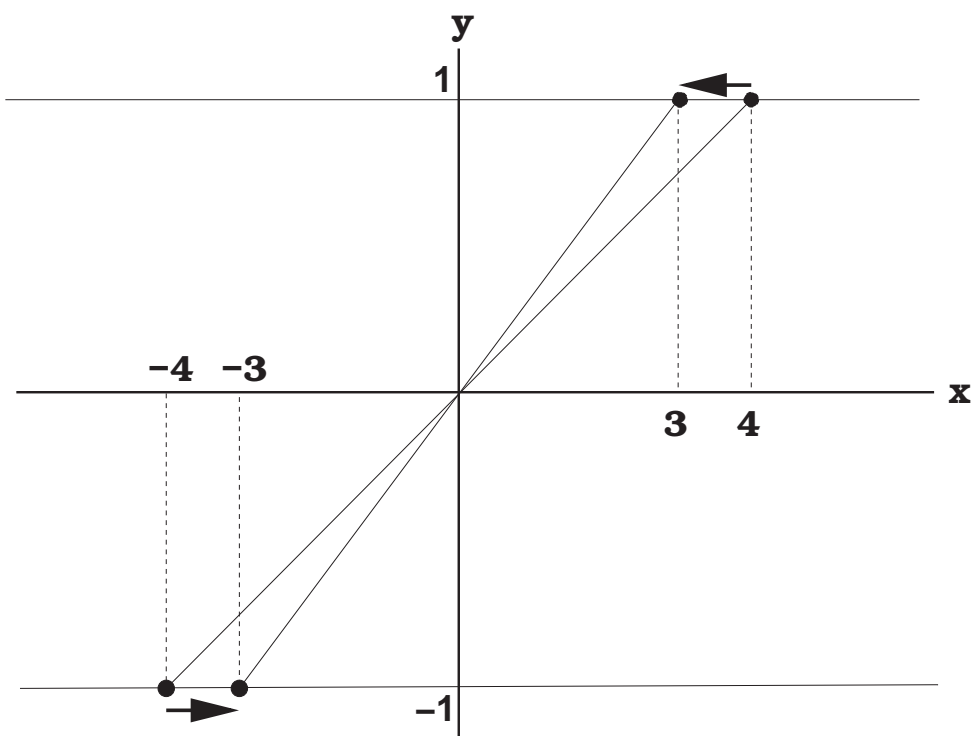


Figure 1: Bias in MPLE with the zero external field.

In the MPLE, two often-used remedies of the boundary problem are the free boundary assumption and omitting the boundary nodes in the analysis, denoted by INNER. However, both procedures have shortcomings as we now discuss. First, the observations on the boundary pixels can be a considerable portion of the entire data set. For example, when the system is a lattice of size $n \times n$, the number of pixels on the boundary is $4n - 4$ and its proportion in the entire system is approximately $4/n$. Accordingly, for small to moderate sized systems, we have a considerable amount of information loss and the variance of parameter estimates is expected to be larger when ignoring the boundary observations. Second, when the external field is assumed to be 0, $\sum_{k \in N(i)} Z_k$ for any $i \in \partial D$ often shrinks toward 0 and the MPLE of β , which is the regression coefficient between $\mathbf{y} = (Z_i + 1)/2$ and $\mathbf{x} = \sum_{k \in N(i)} Z_k$, tends to overestimate the true β when it is positive, as will typically be the case in practice. Figure 1 illustrates how the free boundary condition results in an overestimation of β .

3 Augmented Neighborhood Variable to Adjust the Edge Effect

For the boundary problem addressed in Section 2, we remove those strong assumptions on the external field and treat the entire outer field as *one unobserved random variable* to be estimated from the data. To be specific, let \mathbf{e} denote a latent random variable associated with the external field, taking values -1 or 1 , which is in a neighborhood of all edge sites (see Figure 2).

Motivated by the definition of the conditional auto-logistic model in (1), given the

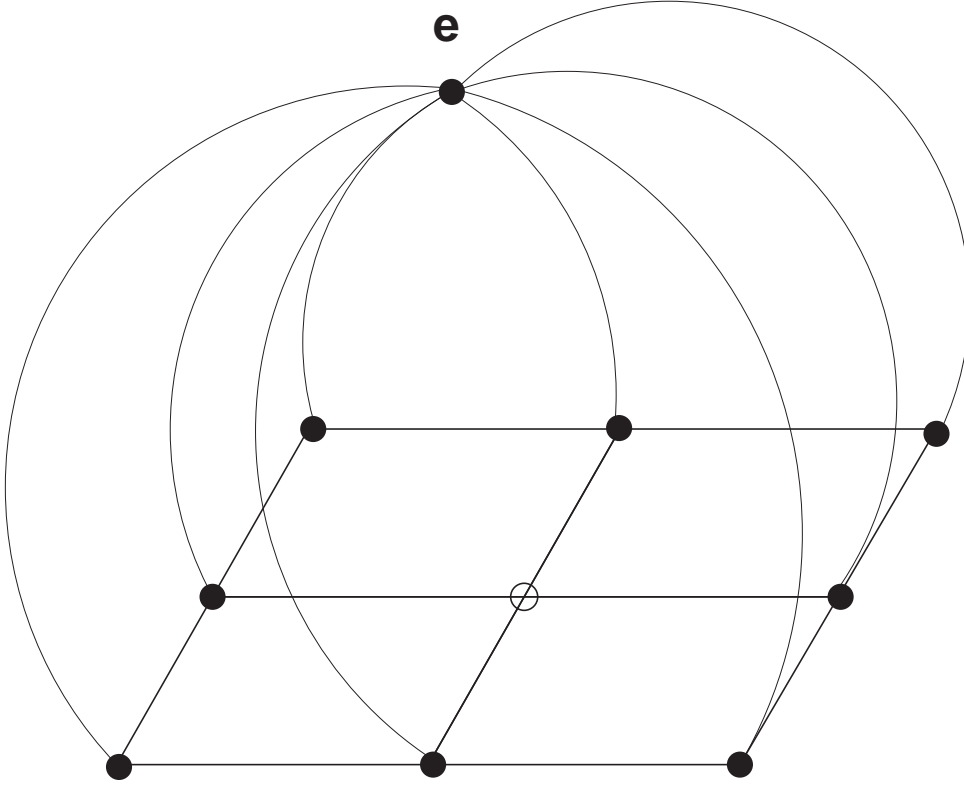


Figure 2: An example of a closed spatial system with the external field. Here, \bullet are boundary points denoted by ∂D .

observations $z = (z_i)^T, i \in D$, the conditional distribution of \mathbf{e} is

$$\mathbb{P}(\mathbf{e} = e | Z = z) = \frac{\exp \{e(\alpha + \beta \sum_{j \in \partial D} z_j)\}}{\exp \{\alpha + \beta \sum_{j \in \partial D} z_j\} + \exp \{-\alpha - \beta \sum_{j \in \partial D} z_j\}}, \quad (5)$$

and the conditional expectation of \mathbf{e} becomes

$$\mathbb{E}(\mathbf{e} | Z = z) = \frac{\exp(2\alpha + 2\beta \sum_{j \in \partial D} z_j) - 1}{\exp(2\alpha + 2\beta \sum_{j \in \partial D} z_j) + 1}. \quad (6)$$

In what follows, we derive the procedures of estimating α and β with the REF, using the technique of Expectation Maximization (EM) for incomplete data.

3.1 MLE with the Augmented Neighborhood Variable

We begin by considering the ML method. From (2), the complete log-likelihood function is

$$\text{LL}(\alpha, \beta; \mathbf{Z}, \mathbf{e}) = \alpha \sum_{i \in D} Z_i + \frac{\beta}{2} \sum_{i \in D} Z_i \sum_{j \in N(i)} Z_j \Big|_{\mathbf{e}} - \log \Psi(\alpha, \beta, \mathbf{e}) \quad (7)$$

where $Z_j|_{\mathbf{e}}$ implies $Z_j = \mathbf{e}$ for $j \notin D$; similarly, $\Psi(\alpha, \beta, \mathbf{e})$ denotes $\Psi(\alpha, \beta)$ defined in (3) with all $Z_j = \mathbf{e}$ for $j \notin D$. Given a realization \mathbf{z} of the random vector \mathbf{Z} , the expected log-likelihood function with respect to the missing component \mathbf{e} becomes

$$\begin{aligned} \text{ELL}(\alpha, \beta) &= \mathbb{E} \left(\log L(\alpha, \beta; \mathbf{Z}, \mathbf{e}) \Big| \mathbf{Z} = \mathbf{z} \right) \\ &= \alpha \sum_{i \in D} z_i + \frac{\beta}{2} \sum_{i \in D} z_i \sum_{j \in N(i)} z_j \Big|_{\mathbf{e}=1} \cdot \mathbb{P}(\mathbf{e} = 1 | \mathbf{Z} = \mathbf{z}) \\ &\quad + \frac{\beta}{2} \sum_{i \in D} z_i \sum_{j \in N(i)} z_j \Big|_{\mathbf{e}=-1} \cdot \mathbb{P}(\mathbf{e} = -1 | \mathbf{Z} = \mathbf{z}) \\ &\quad - \log \Psi(\alpha, \beta, \mathbf{e} = 1) \cdot \mathbb{P}(\mathbf{e} = 1 | \mathbf{Z} = \mathbf{z}) - \log \Psi(\alpha, \beta, \mathbf{e} = -1) \cdot \mathbb{P}(\mathbf{e} = -1 | \mathbf{Z} = \mathbf{z}). \end{aligned} \quad (8)$$

In a typical ‘‘expectation’’ step of the EM technique, $\mathbb{P}(\mathbf{e} = e | \mathbf{Z} = \mathbf{z})$ in (8) is substituted for by (6) estimated at the current value of α and β . So, by differentiating $\text{ELL}(\alpha, \beta)$, the MLEs of α and β are the solution to

$$\begin{aligned} \sum_{k \in D} z_k &= \mathbb{E}_{\alpha, \beta} \left(\sum_{k \in D} Z_k \Big|_{\mathbf{e}=1} \right) \mathbb{P}(\mathbf{e} = 1 | \mathbf{Z} = \mathbf{z}) \\ &\quad + \mathbb{E}_{\alpha, \beta} \left(\sum_{k \in D} Z_k \Big|_{\mathbf{e}=-1} \right) \mathbb{P}(\mathbf{e} = -1 | \mathbf{Z} = \mathbf{z}) \end{aligned} \quad (9)$$

$$\begin{aligned} \sum_{k \in D} z_k \left(\sum_{l \in N_k} z_l \right) \Big|_{\mathbf{e}=1} \mathbb{P}(\mathbf{e} = 1 | \mathbf{Z} = \mathbf{z}) &+ \sum_{k \in D} z_k \left(\sum_{l \in N_k} z_l \right) \Big|_{\mathbf{e}=-1} \mathbb{P}(\mathbf{e} = -1 | \mathbf{Z} = \mathbf{z}) \\ &= \mathbb{E}_{\alpha, \beta} \left(\sum_{k \in D} Z_k \sum_{l \in N_k} Z_l \Big|_{\mathbf{e}=1} \right) \mathbb{P}(\mathbf{e} = 1 | \mathbf{Z} = \mathbf{z}) \\ &\quad + \mathbb{E}_{\alpha, \beta} \left(\sum_{k \in D} Z_k \sum_{l \in N_k} Z_l \Big|_{\mathbf{e}=-1} \right) \mathbb{P}(\mathbf{e} = -1 | \mathbf{Z} = \mathbf{z}). \end{aligned} \quad (10)$$

We approximate the above estimating equation, given the current estimate of \mathbf{e} , $\widehat{\mathbf{e}} = \mathbb{E}(\mathbf{e}|Z = z)$, by

$$\sum_{k \in D} z_k = \mathbb{E}_{\alpha, \beta} \left(\sum_{k \in D} Z_k \mid \widehat{\mathbf{e}} \right) \quad (11)$$

$$\sum_{k \in D} z_k \left(\sum_{l \in N_k} z_l \right) \Big|_{\widehat{\mathbf{e}}} = \mathbb{E}_{\alpha, \beta} \left(\sum_{k \in D} Z_k \sum_{l \in N_k} Z_l \mid \widehat{\mathbf{e}} \right), \quad (12)$$

where the expectation $\mathbb{E}_{\alpha, \beta}$ is w.r.t. Z . Hence, the overall suggested procedure is an iterative procedure between (i) estimate $\widehat{\mathbf{e}} = \mathbb{E}(\mathbf{e}|Z = z)$ using the current estimate of α and β , and (ii) update the current estimate of α and β , say $\widehat{\alpha}^{(k-1)}$ and $\widehat{\beta}^{(k-1)}$, using the current estimate of \mathbf{e} , say $\widehat{\mathbf{e}}$, as

$$\begin{pmatrix} \widehat{\alpha}^{(k)} \\ \widehat{\beta}^{(k)} \end{pmatrix} = \begin{pmatrix} \widehat{\alpha}^{(k-1)} \\ \widehat{\beta}^{(k-1)} \end{pmatrix} - \begin{pmatrix} \lambda_{k1} \\ \lambda_{k2} \end{pmatrix} \cdot \begin{pmatrix} \sum_{k \in D} z_k - \mathbb{E}_{\alpha, \beta} \left(\sum_{k \in D} Z_k \mid \widehat{\mathbf{e}} \right) \\ \sum_{k \in D} z_k \left(\sum_{l \in N_k} z_l \right) \Big|_{\widehat{\mathbf{e}}} - \mathbb{E}_{\alpha, \beta} \left(\sum_{k \in D} Z_k \sum_{l \in N_k} Z_l \mid \widehat{\mathbf{e}} \right) \end{pmatrix}, \quad (13)$$

where λ_{k1} and λ_{k2} is an appropriately chosen decreasing sequence.

3.2 MPLE with the Augmented Neighborhood Variable

We now proceed to consider the MPL method as an alternative to the above ML method for estimating α and β with the augmented neighborhood variable. It is much simpler in computation, which makes it attractive in practice. It is not difficult to see that the complete log-PL can be expressed as

$$\text{LPL}(\alpha, \beta; Z, \mathbf{e}) = \sum_{i \in D - \partial D} \log \mathbb{P}(Z_i | Z_{-i}) + \sum_{i \in \partial D} \log \mathbb{P}(Z_i | Z_{-i}, \mathbf{e})$$

Given the observations $Z = z$, the corresponding expected log-PL (ELPL) is

$$\begin{aligned}
\text{ELPL}(\alpha, \beta) &= \mathbb{E}\left(\text{LPL}(\alpha, \beta; Z, \mathbf{e}) \mid Z = z\right) \\
&= \text{LPL}(\alpha, \beta; Z = z, \mathbf{e} = 1) \cdot \mathbb{P}(\mathbf{e} = 1 \mid z) + \text{LPL}(\alpha, \beta; Z = z, \mathbf{e} = -1) \cdot \mathbb{P}(\mathbf{e} = -1 \mid z) \\
&= \sum_{i \in D - \partial D} \log \mathbb{P}(z_i \mid z_{-i}) + \sum_{i \in \partial D} \log \mathbb{P}(z_i \mid z_{-i}, \mathbf{e} = 1) \cdot \mathbb{P}(\mathbf{e} = 1 \mid z) \\
&\quad + \sum_{i \in \partial D} \log \mathbb{P}(z_i \mid z_{-i}, \mathbf{e} = -1) \cdot \mathbb{P}(\mathbf{e} = -1 \mid z)
\end{aligned} \tag{14}$$

Further, we can consider the mode approximation to (14) to simplify the estimation procedure

$$\text{ELPL}(\alpha, \beta) \approx \sum_{i \in D - \partial D} \log \mathbb{P}(z_i \mid z_{-i}) + \sum_{i \in \partial D} \log \mathbb{P}(z_i \mid z_{-i}, \mathbf{e} = \mathbb{E}(\mathbf{e} \mid z)). \tag{15}$$

Using this mode approximation, we impute \mathbf{e} by $\mathbb{E}(\mathbf{e} \mid z)$ based on the current parameter values and then maximization of (15) can be simply done by running a logistic regression with $Y_i = (Z_i + 1)/2$ against $\sum_{k \in N(i)} Z_k \mid_{\mathbf{e}}$. Alternatively, we can also maximize (14) using a weighted logistic regression with appropriate weights; the weights are 1's for the internal observations, $\mathbb{P}(\mathbf{e} = 1 \mid z)$ for the boundary observations with $\mathbf{e} = 1$, and $\mathbb{P}(\mathbf{e} = -1 \mid z)$ for the boundary observations with $\mathbf{e} = -1$.

Before we end this section, we note that, to obtain the MPLE with the augmented variable, the proposed estimation procedures are iterative between the step of estimating the missing external field and the step of updating the parameter estimates.

4 A Numerical Comparison with Existing Procedures

In this section, we conducted simulation studies to investigate the performance of the proposed MPLE and MLE using the augmented neighborhood model. We also compare them to the estimates using other commonly-used treatments to the boundary problem.

First, to investigate the performance of the MPLE with the REF, we consider three sets of (α, β) : $(0.0, 0.1)$, $(0.0, 0.3)$, and $(0.3, 0.1)$. For each parameter set, we generate 100 data sets with a grid size of 10×10 . For each data set, we generate 40×40 with free boundary condition and take the middle 10×10 values. The first 40×40 data is generated using the Gibbs sampler with 50,000 iterations of sampling the entire system. Here, we restrict our study to those less correlated spatial systems since the MPLE with 10×10 observations shows poor performance when the system is highly correlated; or more technically, there is a problem of singularity in fitting logistic regression models. We take $\lambda_{k1} = \lambda_{k2} = 0.02/(k + 5)$ where λ_{ki} , $i = 1, 2$, are as defined in (13).

For each data set, we compute the MPLE with three different assumptions on the external field: (1) free boundary condition; (2) omitting all boundary values (INNER); (3) the augmented neighborhood model proposed in Section 3, treating the external field as an unobserved random variable to be estimated. A summary of the obtained estimates are presented in Table 1.

Two observations can be made from Table 1. First, we can see that the MPLE with the free boundary condition overestimated the true β in all cases, which confirmed what we suggested in Section 2.2 (see Figure 2). This phenomenon of overestimation is particularly severe for $\beta = 0.3$ and the MPLE with the augmented neighborhood variable shows much smaller bias than that with the free boundary when β is 0.3. Second, when compared to the MPLE with the INNER, the MPLE with the augmented neighborhood variable shows a similar amount of bias but a consistently smaller variance. Conclusively, the MPLE with the augmented neighborhood variable outperforms the other two existing assumptions.

Our second simulation compares the performance of the MLE with the augmented neighborhood variable to that with the free boundary and the torodial boundary condition. We again generate 100 data sets from model (1) with parameter values $(\alpha, \beta) =$

(α, β)	Method	$\hat{\alpha}$ mean (s.e.)	$\hat{\beta}$ mean (s.e.)
(0.0, 0.1)	FREE	-0.002 (0.083)	0.106 (0.077)
	INNER	-0.002 (0.114)	0.101 (0.095)
	ANV	-0.005 (0.076)	0.098 (0.072)
	ANV-MODE	-0.004 (0.074)	0.106 (0.076)
(0.0, 0.3)	FREE	0.001 (0.071)	0.391 (0.088)
	INNER	-0.010 (0.121)	0.315 (0.110)
	ANV	-0.011 (0.109)	0.298 (0.085)
	ANV-MODE	-0.006 (0.112)	0.307 (0.084)
(0.3, 0.1)	FREE	0.321 (0.150)	0.102 (0.094)
	INNER	0.338 (0.189)	0.089 (0.097)
	ANV	0.291 (0.175)	0.095 (0.092)
	ANV-MODE	0.284 (0.171)	0.099 (0.092)

Table 1: “FREE” means MPLE with the free boundary condition, “INNER” is MPLE with only internal observations, “ANV” is MPLE with the augmented neighborhood variables which solves (14), and “ANV-MODE” is MPLE with the augmented variable which solves (15).

(0.0, 0.1), (0.0, 0.3) and (0.3, 0.1), respectively. The Gibbs sampler is used to approximate the derivatives of the partition function (9) and (10). The number of iterations of Gibbs sampler is chosen as 50,000 so that the statistics $\sum_{k \in D} Z_k$ and $\sum_{k \in D} Z_k \sum_{l \in N(k)} Z_l$ converge in distribution. The convergence is checked using the plot of these statistics (computed with Gibbs samples from each iteration) against to the number of iterations. We refer the reader to Brooks and Roberts (1998) and references therein for more elegant ways to diagnose the convergence of the sampler.

The estimates and their standard errors are reported in Table 2. It is not surprising that the free boundary condition performs well when $\alpha = 0.0$, but not when $\alpha = 0.3$ where the augmented neighborhood variable performs better. For the case $\alpha = 0.0$, the augmented neighborhood model is as good as the free boundary model in terms of both bias and variance. On the other hand, the proposed augmented neighborhood model performs better than the torodial boundary model in all cases.

Our final simulation addresses the case when the data is generated from the fixed boundary condition. We fix the boundary values as 1 and generate 100 data sets of size 10×10 using the Gibbs sampler. We set the parameters as $\alpha = 0.0$ and $\beta = 0.3$. Both the MPLE and the MLE with an augmented neighborhood value are superior to the estimates with other boundary assumptions in both α and β estimation. The results are reported in Table 3. As in the previous simulations, the MLE shows larger bias but smaller variance than the MPLE.

5 An Illustrative Example: *Plantago lanceolata*

In this section, we provide an additional illustration of our proposed MLE and MPLE with the augmented neighborhood model by applying them to the *Plantago lanceolata*

(α, β)	Method	$\hat{\alpha}$ mean (s.e.)	$\hat{\beta}$ mean (s.e.)
(0.0, 0.1)	FREE	-0.003 (0.086)	0.097 (0.070)
	TORUS	-0.003 (0.083)	0.095 (0.069)
	ANV-MODE	-0.004 (0.079)	0.104 (0.074)
(0.0, 0.3)	FREE	-0.001 (0.086)	0.285 (0.066)
	TORUS	0.002 (0.068)	0.270 (0.058)
	ANV-MODE	-0.000 (0.057)	0.285 (0.054)
(0.3, 0.1)	FREE	0.332 (0.143)	0.092 (0.084)
	TORUS	0.321 (0.146)	0.091 (0.082)
	ANV-MODE	0.298 (0.152)	0.092 (0.080)

Table 2: MLE: “ANV-MODE” solves (11) and (12) in Section 3.

data.

Plantago lanceolata is a perennial herb. It grows in a temperate climate and its young leaves are often used as a vegetable. In the data set, the state of Kansas was divided into 105 districts and the data set recorded the distribution of *Plantago lanceolata* in these districts of Kansas. Specifically, if a plant was observed in the k -th district, $Y_k = 1$ is recorded; otherwise, $Y_k = -1$ is recorded. There are 40 edge sites, which is a substantial portion of the 105 districts.

To investigate geographic patterns in the distribution of *Plantago lanceolata* in Kansas, we assumed the auto-logistic model defined in the introduction and employed several different estimation procedures, as listed in Table 4. It should be mentioned that the data is observed from a non-lattice spatial system and the other procedures (*e.g.*, MPLE

Estimator	Method	$\hat{\alpha}$ mean (s.e.)	$\hat{\beta}$ mean (s.e.)
MPLE	FREE	0.120 (0.105)	0.304 (0.101)
	INNER	0.011 (0.154)	0.305 (0.114)
	ANV	0.005 (0.137)	0.298 (0.099)
	ANV-MODE	0.005 (0.137)	0.299 (0.099)
MLE	FREE	0.130 (0.106)	0.266 (0.071)
	TORUS	0.089 (0.120)	0.260 (0.084)
	ANV-MODE	0.036 (0.111)	0.278 (0.084)

Table 3: Estimates for the data generated from the model with fixed boundary. In MPLE, “FREE” means MPLE with the free boundary condition, “INNER” is MPLE with only internal observations, “ANV” is MPLE with the augmented neighborhood variables which solves (14), and “ANV-MODE” is MPLE with the augmented variable which solves (15). In MLE, ‘ANV-MODE’ solves (11) and (12) in Section 3.

Method	α estimate (s.e.)	β estimate (s.e.)
MPLE with FREE	-0.042 (0.223)	0.203 (0.076)
MPLE with INNER	0.053 (0.263)	0.175 (0.096)
MPLE with ANV-MODE	0.043 (0.244)	0.212 (0.084)
MLE with ANV-MODE	-0.002 (0.060)	0.240 (0.029)

Table 4: Analysis of *Plantago lanceolata* data: MPLE-FREE is the MPLE with the assumption of the zero external field; MPLE-INNER is the MPLE with only the internal observations; MPLE with ANV-MODE and MLE with ANV-MODE are the MPLE using the mode approximation and the MLE respectively, both with the augmented neighborhood variable introduced in Section 3.

with the torodial boundary condition) discussed previously are not applicable here. The computed estimates are given in Table 4. The standard errors of the MPLEs are computed using a parametric bootstrap procedure and that of the MLE is computed from the Hessian matrix of the log-likelihood function.

In Table 4, all the results indicate that there may exist significant geographic clustering in the distribution of *Plantago lanceolata*. However, the β estimates vary over a relatively wide range. In particular, it is interesting to see that $\hat{\beta}$'s for both the MPLE with the INNER and the MPLE with the free boundary condition are beyond two standard errors of the MLE with the augmented neighborhood variable. Thus, we conjecture that those estimates perform poorly. In contrast, the MPLE with the augmented neighborhood variable was indeed within two standard errors of the MLE with the same boundary condition.

Finally, it should be remarked that the standard errors of the MPLE in Table 4 are larger than those reported in Section 4 (the simulation studies). Note that the simulation

study is implemented on a 10×10 regular grid, but the example has values on an irregular grid with about 100 data points. The difference in standard errors is conjectured to be largely attributable to the irregularity of the space.

6 Discussion

In this paper, we have proposed the idea of accounting for edge effects for binary spatial data. In particular, we treat the external field as an unobserved random component. Two basic estimators, the MLE and the MPLE, have been developed based on the EM technique to iteratively estimate the missing REF. Through our modest simulation studies, we have shown that using the augmented improves the performance of both the estimators. From our example, we have also seen that using the augmented neighborhood model is not restricted by the regularity of the spatial system and can be used in any non-lattice spatial system. In contrast, TORUS is hard to use for irregular spatial systems.

Our discussion in the main body of the paper is restricted to binary spatial data and the auto-logistic model. However, the idea of the random external field can be extended to other spatial auto-regressive (AR) models in a straightforward way. For example, the CAR Gaussian model is defined by

$$z_i | z_{-i} \sim \text{Gaussian}\left(\mu_i + \sum_{j \in N_i} c_{ij} (z_j - \mu_j), \tau_i^2\right),$$

where $c_{ij}\tau_j^2 = c_{ji}\tau_i^2$ and $c_{ii} = 0$; we assume homoscedastic errors $\tau_i^2 = \tau^2$ for every $i \in D$ and, thus, $c_{ij} = c_{ji}$ for every $i \neq j$ for notational simplicity. The procedure for the MLE with the augmented neighborhood model is exactly the same as that of the auto-logistic model and we only discuss the MPLE with the REF. Analogously to the auto-logistic

model, the expected log pseudo-likelihood becomes

$$\begin{aligned} \text{ELPL}(\alpha, \beta) = & \sum_{i \in D - \partial D} \left\{ -\frac{1}{2\tau^2} (z_i - \alpha - \beta \sum_{j \in N_i} z_j)^2 - \frac{1}{2} \log \tau^2 \right\} \\ & + \sum_{i \in \partial D} \left\{ -\frac{1}{2\tau^2} (z_i - \alpha - \beta \sum_{j \in N_i} z_j)^2 - \frac{1}{2} \log \tau^2 \right\} \Big|_{\mathbf{e} = \mathbf{E}(\mathbf{e} | Z = \mathbf{z})}. \end{aligned} \quad (16)$$

Thus, we can evaluate the MLE using the following EM procedure: (i) given the current estimates of α and β , compute $\mathbf{e} = \mathbf{E}(\mathbf{e} | Z = \mathbf{z})$; (ii) maximize (16) using the (weighted) regression between z_i and $\sum_{j \in N_i} z_j$ as in the auto-logistic model.

Acknowledgment

The *Plantago lanceolata* data is from the PLANTS Database (<http://plants.usda.gov>), National Plant Data Center, Baton Rouge, LA 70874-4490 USA.

References

- Anselin, L. and Cho, W. 2002, Spatial Effects and Ecological Inference, *Political Analysis*, 10, 276-297.
- Bartolucci, F. and Besag, J. 2002, A Recursive Algorithm for Markov Random Fields, *Biometrika*, 89, 724-730.
- Besag, J. 1974, Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society-B*, 23, 192-236.
- Besag, J. 1975, Statistical Analysis of Non-lattice Data, *The Statistician*, 24, 179-195.

- Besag, J. 1986, On the Statistical Analysis of Dirty Pictures, *Journal of the Royal Statistical Society-B*, 48, 259-302.
- Brooks, S.P. and Roberts, G.O. 1998, Assessing Convergence of Markov Chain Monte Carlo Algorithms, *Statistics and Computing*, 8, 319-335.
- Comets, F. 1992, On Consistency of a Class of Estimators for Exponential Families of Markov Random Fields on the Lattice, *The Annals of Statistics*, 20, 455-468.
- Cressie, N.A.C. 1993. *Statistics for Spatial Data*, Wiley, New York.
- Dempster, A., Laird, N., and Rubin, D. 1977, Maximum Likelihood from Incomplete Data Via the EM algorithm, *Journal of the Royal Statistical Society-B*, 39, 1-38.
- Geman, S. and Geman, D. 1984, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- Geman, S. and Graffigne, C. 1986, Markov Random Field Image Models and Their Applications to Computer Vision, *Proceedings of the International Congress of Mathematicians*, Berkeley, CA.
- Geyer, C.J. and Thompson, E.A. 1992, Constrained Monte Carlo Maximum Likelihood For Dependent Data, *Journal of the Royal Statistical Society-B*, 54, 657-699.
- Guyon, X. and Künsch, H.R. 1992, Asymptotic Comparison of Estimators in the Ising Model, *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, *Lecture Notes in Statistics*, 74, 177-198: Springer, Berlin.

- 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.
- Huang, F. and Ogata, Y. 2002, Generalized Pseudo-Likelihood Estimates for Markov Random Fields on Lattice, *Annals of the Institute of Statistical Mathematics*, 54, 1-18.
- Moyeed, R.A. and Baddeley, A.J. 1991, Stochastic Approximation of the MLE for a Spatial Point Pattern, *Scandinavian Journal of Statistics*, 18, pp. 39-50.
- Pickard, D.K. 1982, Inference for General Ising Models, *Journal of Applied Probability*, 19A, *Essays in Statistical Sciences: Papers in Honour of P.A.P. Moran*, 345-357.
- Asymptotic Inference for an Ising Lattice, *Journal of Applied Probability*, 13, pp. 486-497.
- Pickard, D.K. 1987, Inference for Discrete Markov Fields: The Simplest Nontrivial Case, *Journal of the American Statistical Association*, 82, 90-96.
- Possolo, A. 1986, Estimation of Binary Markov Random Fields, Technical Report no 77. Department of Statistics, University of Washington.
- Revees, R. and Pettitt, A.N. 2004, Efficient Recursion for General Factorisable Models, *Biometrika*, 91, 751-757.
- Sherman, M., Apanasovich, T. and Carroll, R.J. 2003, On Estimation in Binary Auto-logistic Spatial Models, to appear *Journal of Statistical Computation and Simulation*.
- Younes, L. 1991, Maximum Likelihood Estimation for Gibbs Fields, *Spatial Statistics and Imaging* (ed. A. Possolo), *Lecture Notes, Monograph Series*, 20, 403-426, Institute of Mathematical Statistics, Hayward, California.

