

On Estimation in Binary Autologistic Spatial Models

Michael Sherman ¹, Tatiyana V. Apanasovich ², and Raymond J. Carroll ^{3 4}

Abstract

There is a large and increasing literature in methods of estimation for spatial data with binary responses. The goal of this article is to describe some of these methods for the autologistic spatial model, and to discuss computational issues associated with them. The main way we do this is via illustration using a spatial epidemiology data set involving liver cancer. We first demonstrate why Maximum Likelihood is not currently feasible as a method of estimation in the spatial setting with binary data using the autologistic model. We then discuss alternative methods, including Pseudo Likelihood, Generalized Pseudo Likelihood, and Monte Carlo Maximum Likelihood estimators. We describe their asymptotic efficiencies and the computational effort required to compute them. These three methods are applied to the data set and compared in a simulation experiment.

Keywords: Autologistic, Bootstrap, Coding, Maximum Likelihood, Monte Carlo Maximum Likelihood, Pseudo Likelihood, Resampling, Spatial models.

Short Title: Binary Autologistic Spatial Models

¹(E-mail: sherman@stat.tamu.edu), Department of Statistics, TAMU 3143, Texas A&M University, College Station TX 77843–3143, USA.

²(E-mail: tanya@stat.tamu.edu), Department of Statistics, TAMU 3143, Texas A&M University, College Station TX 77843–3143, USA.

³(E-mail: carroll@stat.tamu.edu), Department of Statistics, TAMU 3143, Texas A&M University, College Station TX 77843–3143, USA.

⁴The research of the second and third authors was supported by a grant from the National Cancer Institute (CA-57030). Research of all three authors supported by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

1 INTRODUCTION

There is a large and increasing literature in methods of estimation for spatial data with binary responses. The goal of this article is to describe some of these methods for the autologistic spatial model, and to discuss computational issues associated with them. The main way we do this is via illustration using a spatial epidemiology data set involving liver cancer. We first demonstrate why Maximum Likelihood is not currently feasible as a method of estimation in the spatial setting with binary data using the autologistic model. We then discuss alternative methods, including Pseudo Likelihood, Generalized Pseudo Likelihood, and Monte Carlo Maximum Likelihood estimators. We describe their asymptotic efficiencies and the computational effort required to compute them.

Many of the uses of the autologistic spatial model involve spatial epidemiology applications. One celebrated example of spatial epidemiology is John Snow's (1855) study of the London cholera epidemic. United States cancer mortality maps have been compiled by Riggan et al. (1987) for use in "developing and examining hypotheses about the influence of various environmental factors" (p. xi) and for investigating possible associations of cancer with "unusual demographic, environmental, industrial characteristics, or employment patterns (p. xii). Consider cancer of the liver and gallbladder (including bile ducts) for white males during the decade from 1950-1959, see Figure 1. There is some apparent geographic clustering. The question is whether this is a significant feature of the data, statistically and otherwise.

This paper is organized as follows. Section 2 describes maximum likelihood estimation for the autologistic spatial model, especially the difficulty in computing the estimator for large samples. We conclude that at present, there is no simple way to compute the likelihood function itself, and hence to make likelihood inference. Section 3 describes three alternatives to maximum likelihood: Pseudolikelihood, Generalize Pseudolikelihood and Monte-Carlo Maximum Likelihood. As we describe each method, we illustrate their application using the liver cancer data set, as well as in a simulation study. Section 4 has concluding remarks.

2 MAXIMUM LIKELIHOOD

In this section we demonstrate why maximum likelihood cannot be applied in the usual way for the autologistic model for binary spatial data. As will be seen, this is true even for moderately sized data sets, and the principle holds for many other spatial data structures.

Let $\{Z_i : i \in D\}$ denote our observed data, where D denotes the set of indices where we observe our data and let $|D|$ denotes the total number of observations. For each location i , $N(i)$ denotes a set of “neighbors”. This is defined explicitly in each particular setting and typically depends on distance. The observations may be continuous or discrete. We focus on binary outcomes, although the issues addressed apply more generally.

Let D denote the set of indices in Figure 1, which has 2003 locations so that $|D| = 2003$. For each location i let the neighborhood $N(i)$ denote the four nearest-neighbor indices of i . Code $Z_i = +1$ if the cancer mortality rate is “high” at site i , and code $Z_i = -1$ if the rate is “low”. This was done using the quantiles of the observed rates from the maps in Riggan et al. and classifies approximately 27% of all counties as having a high rate.

Assume the data were generated by an autologistic model (Besag, 1974),

$$P[Z_i = z_i | Z_j = z_j : j \neq i] = \frac{\exp\{z_i(\alpha + \beta \sum_{j \in N(i)} z_j)\}}{\exp(\alpha + \beta \sum_{j \in N(i)} z_j) + \exp(-\alpha - \beta \sum_{j \in N(i)} z_j)}, \quad (2.1)$$

so that the conditional distribution of Z_i depends on all other observations only through the sum of its four nearest neighbors. The parameter α determines the overall proportion of high mortality rates while the parameter β determines the strength of clustering in the data. Of course, $\beta = 0$ corresponds to no clustering or no dependence of a county’s rate with neighboring county rates.

Given this model, the goal becomes to estimate the parameters β and α and draw inferences on these parameters. Due to its asymptotic optimality in a variety of settings it is natural to attempt to fit and assess accuracy using maximum likelihood. It can be shown (e.g., Cressie, 1993) that the joint likelihood of the data is

$$L(\alpha, \beta) := \frac{\exp\{\alpha \sum_{i \in D} Z_i + (\beta/2) \sum_{i \in D} Z_i (\sum_{j \in N(i)} Z_j)\}}{\sum_{z_1, \dots, z_n} \exp\{\alpha \sum_{k \in D} z_k + (\beta/2) \sum_{k \in D} z_k (\sum_{\ell \in N(k)} z_\ell)\}}, \quad (2.2)$$

where \sum_{z_1, \dots, z_n} denotes the sum over all $2^{|D|}$ possible realizations. For our data set the denominator has 2^{2003} summands. Thus for a larger number of covariates, or for count or

continuous responses this problem becomes even worse. This example shows why it is not, in general, practical to carry out maximum likelihood estimation in the spatial setting by direct enumeration.

The failure in the attempt to apply maximum likelihood by direct enumeration leads to the search for alternative methods of estimation. A natural approach is to try to approximate the normalizing constant in the denominator of the likelihood. Let $\{z_i^{(1)}\}, i = 1, \dots, 2003$, be an independent random sample of variables from the Bernoulli distribution. Then each $z_i^{(1)}$ is equal to 1 with probability 0.5 and -1 with probability 0.5. We can generate B data sets in this fashion and use

$$B^{-1} \sum_{b=1}^B [\exp\{\alpha \sum_{k \in D_n} z_k^{(b)} + (\beta/2) \sum_{k \in D_n} z_k^{(b)} (\sum_{\ell \in N(k)} z_\ell^{(b)})\}] \quad (2.3)$$

to approximate

$$\sum_{z_1, \dots, z_n} [\exp\{\alpha \sum_{k \in D} z_k + (\beta/2) \sum_{k \in D} z_k (\sum_{\ell \in N(k)} z_\ell)\}] / 2^n.$$

This approximation is consistent for the normalizing constant as B approaches ∞ . Noting that maximum likelihood estimators do not depend on any constant multiple suggests using $\sum_{b=1}^B \exp\{\alpha \sum_{k \in D} z_k^{(b)} + \beta \sum_{i \in D} z_i^{(b)} (\sum_{\ell \in N(i)} z_\ell^{(b)})\}$ in place of the denominator of $L(\alpha, \beta)$.

In this data set $\sum_{i \in D} Z_i = -923$ and $\sum_{i \in D} Z_i (\sum_{j \in N(i)} Z_j) = 2326$ so we seek to maximize:

$$\exp(-923\alpha + 1163\beta) / \sum_{b=1}^B \exp\{\alpha \sum_{k \in D} z_k^{(b)} + (\beta/2) \sum_{k \in D} z_k^{(b)} (\sum_{\ell \in N(k)} z_\ell^{(b)})\}.$$

Initially, this procedure appears reasonable, but it does not work well in practice. To appreciate why, note that each sum $\sum_{k \in D} z_k^{(b)}$, $b = 1, \dots, B$ is approximately normally distributed with mean 0 and variance $|D|$. Thus, the probability that we observe a value -923 or smaller is approximately $P[Z < -923/2003^{1/2}] \simeq 0$. Hence, even for large values of B , e.g., $B = 1,000,000$, it's likely that no term in the denominator sums $\sum_{k \in D} z_k^{(b)}$ is comparable in magnitude to -923 and thus we are very likely to obtain $\alpha = -\infty$. Arguing similarly with regards to β we see that $\alpha = -\infty$ and $\beta = \infty$ will maximize this approximation to the likelihood with very high probability.

The problem is that we have not sampled from any outcomes that are similar to our observed outcome. Thus, it seems natural to consider B stratified outcomes corresponding

to a broad range of possible outcomes. Consider the following outcomes: 1) approximately half the observations have a high rate and half have a low cancer rate, 2) all observations have a low rate and 3) all observations have a high rate. These roughly correspond to $\sum_{i \in D} Z_i = 0$ and $\sum_{k \in D} z_k(b)(\sum_{\ell \in N(k)} z_\ell(b))=0$, $\sum_{i \in D} Z_i = -2003$ and $\sum_{i \in D} Z_i \{\sum_{\ell \in N(k)} z_\ell(b)\} = 8012$, and $\sum_{i \in D} Z_i = 2003$ and $\sum_{i \in D} Z_i \{\sum_{\ell \in N(k)} z_\ell(b)\} = 8012$, respectively. Considering just these three cases we see that one of these three terms will completely dominate the numerator and thus $\alpha = 0$ and $\beta = 0$ will maximize the corresponding approximate likelihood. Thus, this procedure also does not work very well.

We see that we need to weight each of the B stratified outcomes by w_b , the approximate probability of observing this outcome. This procedure works in principle but leads to impractical arithmetic. For example, the denominator summand corresponding to $\sum_{i \in D} Z_i = 2003$ has a very large influence in estimation of α (as discussed in the previous paragraph) but has a weight corresponding to the probability of a z-score of $(2003 - 0)/(2003)^{1/2} = (2003)^{1/2} = 44.8$. The product of the weight and the term is well behaved in theory, but there appears to be no practical way to carry out this computation.

This discussion shows that there is apparently no simple way to carry out maximum likelihood estimation for the autologistic model for spatial data. While it may be possible to do this, no current method can circumvent the computational difficulties that we have discussed. In the next section we describe three feasible alternatives to maximum likelihood estimation.

3 ALTERNATIVES TO MAXIMUM LIKELIHOOD

3.1 Coding and Pseudo Likelihood

Besag (1974), noted that $P(Z_i = z_i | Z_j : j \neq i)$ and $P(Z_k = z_k | Z_j : j \neq k)$ are conditionally independent if $i \notin N(k)$ and $k \notin N(i)$. For the cancer data with 4 nearest neighbors this corresponds to $d(i, k) > 1$. This means that 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 data in each group applies. This gives

two sets of estimates which can be combined (by averaging) to form one estimator.

Instead of combining the coding estimators it appears reasonable to simply pool all components together to form a single likelihood. The resulting “pseudo likelihood” (Besag, 1975) is

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

The corresponding maximum pseudo likelihood estimators (MPLE) are the parameter values that maximize $PL(\alpha, \beta)$. The large sample consistency and asymptotic normality of MPLE’s have been demonstrated by, e.g., Geman and Graffigne (1986), Comets (1992) and Guyon and Künsch (1992), and the method has been applied in a variety of settings.

For the cancer of the liver and gallbladder data set the MPLE’s are $\hat{\alpha} = -0.3224$ and $\hat{\beta} = 0.1056$. The $\log\{PL(\alpha, \beta)\}$ surface is given in Figure 2. The parameter values indicate an overall majority of low rate counties ($\hat{\alpha} < 0$) and that there is some indication of clumping or positive correlation between neighboring counties ($\hat{\beta} > 0$).

The fundamental statistical question is whether the value $\hat{\beta} = 0.1056$ indicates a significant tendency for high cancer rate counties to cluster together ($\beta > 0$). In order to answer this question we require an estimate of the standard error of $\hat{\beta}$.

Standard errors for the PL usually have no closed form solution so it is a nontrivial problem to estimate them. Roughly, there are model free and model based resampling approaches. The model free approach uses block resampling to preserve the correlation structure, while the model based approach resamples from the estimated model. There is a large literature in model free resampling/subsampling for temporally or spatially correlated data; see, for example, Hall (1988), Künsch (1989), Liu and Singh (1992), Lahiri (1991), Politis and Romano (1994), Sherman and Carlstein (1994) and Heagerty and Lumley (2000). The model based approach is asymptotically more efficient under the correct model than the model free approach but is less efficient when the model is not correctly specified. Further, very large data sets may be necessary to realize the increased efficiency of the model based approach, see, e.g., Sherman (1998) for details in the case of correlated time series data.

As an example of the model free approach, the counties in the cancer data set are split up into ten, partially overlapping, subshapes D_k , $k = 1, \dots, 10$. Each subshape is identical

in size ($|D_k| = 141$) and similar in shape to the original data domain. The estimates of model parameters are found on each subshape, yielding “replicate” statistics $\hat{\alpha}^k$ and $\hat{\beta}^k$, $k = 1, \dots, 10$. Then, for example, the estimate of $\text{Var}(\hat{\beta})$ is

$$\frac{|D_k|}{|D|} \sum_{k=1}^{10} \frac{(\hat{\beta}^k - \bar{\beta})^2}{10},$$

where $\bar{\beta} = \sum_{k=1}^{10} \hat{\beta}^k / 10$. This is simply the empirical variance of the standardized replicates, $\hat{\beta}^k$, $k = 1, \dots, 10$. The variability of $\hat{\alpha}$ is estimated analogously. The square roots of these quantities give the resulting estimates of $s.e.(\hat{\alpha}) = .044$ and $s.e.(\hat{\beta}) = .026$. See Sherman and Carlstein (1994) for further details and large sample justification of this method.

A model based approach to variance estimation generates simulated data sets from the estimated model (2.1), computes the estimate from each and then computes the sample variance of the estimates. The creation of the simulated data sets can be accomplished using the Gibbs Sampler. Specifically, we want to generate B simulated data sets from model (2.1) using the estimated parameters from PL. Then we use the sample variance of the B parameter estimates to estimate the variability of the PL estimators.

We need a starting point from which to generate the simulated data sets. Consider independent binary variables from $[-1, 1]$ at each location and let $\mathbf{u}^{(0)} = \{u_i^{(0)}, i = 1, \dots, |D|\}$ denote this initial state.

We update each observation u_i according to (2.1) and the values of its four nearest neighbors. After updating each location we have the “data” $\mathbf{u}^{(1)} = \{u_i^{(1)}\}$, say. Let $\mathbf{u}^{(k)} = \{u_i^{(k)}\}$ denote the observations after k steps of the Gibbs sampler. We may not want to simply take all the $\mathbf{u}^{(k)}$ ’s as realizations from the model for two reasons. First, we started from an independent field whose distribution is not close to the desired joint distribution. Thus, we may want to wait before accepting a given $\mathbf{u}^{(k)}$ as the first pseudo realization. Further, at the k and $k + 1$ steps the observations $\{u_i^k\}$ and $\{u_i^{k+1}\}$ may be heavily correlated. These considerations lead to defining our Monte Carlo sample to be: $\mathbf{z}^{(b)} = \mathbf{u}^{(d+b\gamma)}$, $b = 1, \dots, B$, where d and γ are integers. We see that there are 3 choices of tuning parameters to be made in implementing the Gibbs Sampler:

- The number of Gibbs sampler steps until we accept the $b = 1$ simulated data set (the

“burn in” d);

- The number of Gibbs sampler steps taken between accepted simulated data sets b and $b + 1$ (the “spacing” γ)
- The total number of simulated data sets (B).

The total number of Gibbs steps is $n = d + B\gamma$.

In general, the variance estimates depend on these three choices. Note, in particular, that the choices of γ and B are closely related for fixed n . Increasing the spacing leads to more independent simulated data sets, but decreases the number of simulated data sets. Table 1 gives the standard errors for a variety of choices of burn ins, spacing, and numbers of simulated data sets. The estimates agree reasonably well across all choices of the three tuning parameters. Summarizing, we obtain $s.e.(\hat{\alpha}) \simeq .036$ and that the $s.e.(\hat{\beta}) \simeq .019$. Thus, using either the model based or the model free estimates of variability we conclude that the observed geographic clustering is indeed significant. This conclusion is corroborated by the results of a goodness-of-fit analysis as in Section 7.1 of Besag (1974).

The PL estimators can be obtained using standard software for logistic regression with response variables Z_i and corresponding covariate $\sum_{j \in N(i)} Z_j$. This is an attractive property of these estimators. We stress, however, that the reported standard errors will be wrong as they erroneously assume independence between terms in the PL.

3.2 Generalized Pseudo Likelihood (GPL)

As discussed, it is relatively easy to find the maximizer of the PL. However, the PL is not the true likelihood of the data and thus the estimators may not be efficient relative to the best possible. However, when $\beta = 0$, PL is asymptotically equivalent to ML. (e.g., Cressie (1993)). Thus, we expect that if the correlation between observations is relatively weak (i.e., β small), then the PL method should give relatively efficient estimates.

For example, in the autologistic model with $\alpha = 0$, Huang and Ogata (2002) show via simulation that $\hat{\beta}_{PL}$ has efficiencies of .967, .928, .805, and .563 relative to maximum likelihood for $\beta = .1, .2, .3, .4$ for data on a 64×64 grid. Thus if the true $\beta \simeq .1$ in the cancer

data set there is no great loss in using the computationally efficient PL estimator. We note that $\beta \geq .44$ corresponds to very strong (long range) correlation, see, e.g., Pickard (1987).

Due to the low efficiencies of PL for heavily correlated observations, Huang and Ogata propose Generalized Pseudo-Likelihood (GPL). They observe that the maximum likelihood estimators can be computed for data sets of sizes less than 9 – 15 say, since the denominator in (2.2) has between $2^9 = 512$ and $2^{15} = 32,768$ terms. The proposal is to form the product of joint likelihoods over small subsets of the dataset. We take each of these subsets to be of the same size. Specifically, let $g(i)$ denote a set of points adjacent to (and including) point i and let $\tilde{Z}_{g(i)} = \{Z_k : k \in g(i)\}$ denote the observed values in this group of points. Denote the complement sites to those in $g(i)$ by $\tilde{Z}^{g(i)}$. Then $GPL(\alpha, \beta)$ is (for constant $|g(i)| = g$) $\prod_{i \in D} f(\tilde{Z}_{g(i)} | \tilde{Z}^{g(i)})$, where in our nearest neighbor model

$$f(\tilde{Z}_{g(i)} | \tilde{Z}^{g(i)}) = \frac{\exp\{\alpha \sum_{k \in g(i)} Z_k + \beta \sum_{k \in g(i)} Z_k (\sum_{j \in N(k)} Z_j)\}}{\sum_{z_k: k \in g(i)} \exp\{\alpha \sum_{k \in g(i)} z_k + \beta \sum_{k \in g(i)} z_k (\sum_{j \in N(k)} z_j)\}}.$$

Note that when $g(i) = \{i\}$ the GPL reduces to the PL. Thus GPL can be seen as a compromise between the PL and ML in terms of computational intensity. Huang and Ogata show that this is also true in terms of efficiency of estimation. It is natural to expect that the efficiency of GPL should increase as the number of elements in $\tilde{Z}_{g(i)}$, $|\tilde{Z}_{g(i)}|$, increases. This turns out to be the case to some extent. For example, when $\beta = .3$ in the autologistic model, the efficiencies are .805, .913, .973, .967 when $|\tilde{Z}_{g(i)}| = 1, 5, 9$, and 13, respectively. Efficiencies are estimated via simulation. Approximate MLE's are computed for the square grids under a “periodic boundary” assumption which we discuss shortly.

We considered three estimation schemes in the cancer data set: $|\tilde{Z}_{g(i)}| = 5$ corresponds to a point together with its 4 nearest neighbors at distance one, $|\tilde{Z}_{g(i)}| = 9$ includes the diagonals at distances $2^{1/2}$, while $|\tilde{Z}_{g(i)}| = 13$ includes the points at distance 2. The resulting estimators are given in Table 2. The estimators are similar to those from PL.

We have discussed the efficiency advantages of GPL estimation but there is one sense in which it may decrease efficiency. In spatial models the user typically confronts edge effects. The edge effect refers to spatial locations that do not have fully observed neighbors. For example, for observations on an $n \times n$ grid there are $4n - 4$ edge sites: $4n - 8$ have 3 neighbors and 4 sites have only 2 neighbors. The issue is how to treat these sites that have incomplete

information as (2.1) is not defined.

The issue is subtle: the efficiency of GPL slightly decreased when $|\tilde{Z}_{g(i)}|$ increased from 9 to 13 in Huang and Ogata’s experiment. This often happens and may well be due to edge effects, as we now explain. One possibility is to condition on the edge sites. Thus, they appear as neighbors to internal sites, but do not otherwise appear in the likelihood. For the cancer data, the original data set has 2285 observations but only $|D| = 2003$ of those have all four neighbors. This is equivalent to using .877 of the data (directly) in the PL. For GPL, conditioning gives that the g_5 GPL has 1747 terms $f(\tilde{Z}_{g(i)}|\tilde{Z}^{g(i)})$ in the product $\prod_{i \in D} f(\tilde{Z}_{g(i)}|\tilde{Z}^{g(i)})$ while g_9 has 1677 terms and g_{13} has 1545 terms. This loss is partially due to the irregularly shaped domain but we have seen the same phenomenon occurs for data arranged in an $n \times n$ square.

Another possibility, used by Huang and Ogata, is to assume a “periodic boundary” condition which amounts to wrapping an $n \times n$ square on a torus. This allows for use of (2.1) for all locations, but in practice makes, for example, locations $(1, j)$ and (n, j) neighbors, which is not very appealing. Note that the efficiency of GPL slightly decreased when $|\tilde{Z}_{g(i)}|$ increased from 9 to 13 in Huang and Ogata’s experiment. This suggests that this method does not entirely account for edge effects. Further, it is not clear how to wrap an irregularly shaped domain onto a torus.

A third possibility is to integrate out the “missing” neighbors from the conditional distribution (2.1). This should have reasonable numerical properties, but the resulting conditional probability will not be of the form (2.1). At present, there is no simple solution to the problem of edge effects.

3.3 Monte Carlo Maximum Likelihood

In Section 2 we saw that there is no simple method to approximate the joint likelihood of the observations. Several authors, however, have used Monte Carlo methods to approximate the Maximum Likelihood estimator. Such MCMC stochastic algorithms have been proposed by, e.g., Younes (1991), Moyeed and Baddeley (1991), Geyer and Thompson (1992), and Seymour and Ji (1996), Gu and Zhu (2001). We consider the Monte Carlo Maximum Likelihood

approach (MCML) of Geyer and Thompson.

The basic approach is a slight modification on the likelihood approximation approach in Section 2 that failed due to unstable arithmetic. They consider

$$d_n(\alpha, \beta) = B^{-1} \sum_{b=1}^B \exp \left[(\alpha - \alpha_0) \sum_{k \in D} z_k^{(b)} + \{(\beta - \beta_0)/2\} \sum_{k \in D} z_k^{(b)} \left(\sum_{\ell \in N(k)} z_\ell^{(b)} \right) \right],$$

where α_0 and β_0 are initial values for the parameters, and $z_j^{(b)}, j = 1, \dots, |D|$ is the b^{th} realization, $b = 1, \dots, B$ of data from the autologistic model with parameters α_0, β_0 . These realizations can be obtained by an application of the Gibbs Sampler with conditional distributions given by (2.1) as described in Section 3.1.

Then $\log\{L(\alpha, \beta)\} + \log\{\mathcal{G}(\alpha_0, \beta_0)\} = \alpha \sum_{i \in D} Z_i + (\beta/2) \sum_{i \in D} Z_i (\sum_{j \in N(i)} Z_j) - \log\{d(\alpha, \beta)\}$ where $\mathcal{G}(\alpha_0, \beta_0)$ is the denominator of $L(\alpha_0, \beta_0)$ in equation (2.2) and $d(\alpha, \beta) = \mathcal{G}(\alpha, \beta)/\mathcal{G}(\alpha_0, \beta_0)$. The term $d_n(\alpha, \beta)$ approaches $d(\alpha, \beta)$ as $B \rightarrow \infty$. The term $\log\{\mathcal{G}(\alpha_0, \beta_0)\}$ is taken to be a constant and thus the maximizer of $\alpha \sum_{i \in D} Z_i + (\beta/2) \sum_{i \in D} Z_i (\sum_{j \in N(i)} Z_j) - \log\{d_n(\alpha, \beta)\}$ approaches the MLE as $B \rightarrow \infty$. In practice, we find the parameters that maximize $\sum_{b=1}^B \exp[(\alpha - \alpha_0)u_{1,b} + \{(\beta - \beta_0)/2\}u_{2,b}]$ where $u_{1,b} = \sum_{i \in D} z_i^{(b)} - \sum_{i \in D} Z_i$ and $u_{2,b} = \sum_{i \in D} z_i^{(b)} (\sum_{j \in N(i)} z_j^{(b)}) - \sum_{i \in D} Z_i (\sum_{j \in N(i)} Z_j)$. This yields the same parameter estimates and is more computationally stable.

To see the connection between this approach and the one that failed in Section 2 simply take $\alpha_0 = \beta_0 = 0$ and we see that $d_n(\alpha, \beta)$ is then equal to equation (2.3). Thus, the approximant in (2.3) is a special case of the MCML approach. This shows that although $d_n(\alpha, \beta) \rightarrow d(\alpha, \beta)$ as $B \rightarrow \infty$ it may do so very slowly. Good initial values for the parameters are very important for reasonable convergence. Natural candidates are the MPL or one of the MGPL estimators.

Using the PL estimators as the initial values we computed the MCML estimators of model parameters in the cancer data set using a variety of choices of number of Gibbs steps, burn in, and spacing. The resulting estimators and standard errors based on a single Markov chain of length 10,000 are given in Table 3. All standard errors are computed from the Fisher Information from the Monte Carlo Likelihood.

First we note that the estimators are relatively in agreement across all choices of numbers

of simulations, burn in, and spacings. For example, all estimates of $\hat{\alpha}$ and $\hat{\beta}$ are within approximately 1% of one another. Thus, in this example 2,000 simulations is already more than adequate. This is likely due to the strength of correlation in the data. If the true correlation is stronger (large β), then results based on a large number of simulations, a reasonably large burn in, and possibly larger spacing will be superior. We examine the effect of the choices of spacing and number of simulations in our simulation results. The Monte Carlo Likelihood is given in Figure 3 (for $n = 10,000$, $d = 1,000$, $\gamma = 5$).

Assuming that the longest chain gives the most accurate results, we find that $\hat{\alpha} \simeq -0.304$ with an estimated SE of $\simeq .0345$ and that $\hat{\beta} \simeq 0.117$ with an estimated SE of $\simeq 0.0185$. This suggests that the MCML approach gives estimators that differ from the PL estimators by about 6% for α and about 10% for β . The PL estimates are closer to the MCML estimates than are the GPL estimates, but all estimates are within 1.5 standard deviations of each other.

The model based standard errors based on the PL estimators and the MCML estimators are even closer than the parameter estimates, differing by about 3%. There is a slight indication that the MCML estimators are more efficient than the PL estimators (assuming the estimated SE's are correct). On the other hand, the MCML approach is much more computationally intensive than MPL. In this specific example, we suspect that there is no great efficiency increase by using MCML over MPL. This is because the correlation as given by β is significantly larger than 0, but is small enough so that observations separated by relatively short distances become approximately independent. This suggests that in our setting, it may not be worth the extra computational burden to carry out MCML. On the other hand, if one chooses a model-based method in PL estimation to estimate standard errors, we have seen that similar Gibbs sampling is necessary.

3.4 Simulation Study

To compare the PL, GPL, and MCML estimators for more strongly correlated observations we performed the following simulation study. We considered observations generated from model (2.1) for the values of $\beta = 0.33$ and 0.50. In both cases we took $\alpha = 0.0$. The goal

is to see the effect of the strength of correlation (as given by β) on the different methods of estimation. The estimation methods were: 1) PL, 2-4) GPL with g_5, g_9, g_{13} , (GPL5, GPL9, and GPL13) and 5-6) MCML with 5,000 simulations, burn in=100 and spacing either 2 or 5, 7)-8) MCML with 20,000 simulations, burn in=100 and spacing either 10 or 20 and 9-10) MCML with 200,000 simulations and spacing either 20 or 100. The resulting average parameter estimates and standard deviations from 100 simulated data sets are given in Table 4.

Considering the $\beta = 0.33$ case we see that all estimators are approximately unbiased for α , but there is some indication of a small negative bias in estimation of β . These biases are more significant for the PL and GPL methods than for MCML. However, the largest bias from GPL9 accounts for only approximately 12.5% of MSE. In comparing variability we see that the GPL estimates are the least variable and the MCML estimates are the most variable. Thus, although theory suggests that the MCML approximate the (efficient) MLE, careful monitoring of the Markov chain is necessary to get the best performance from the MCML approach. See, e.g., Huffer and Wu (1998) who employ a method to update the initial parameter estimates (α_0, β_0) as the Markov Chain evolves.

The $\beta = 0.5$ case corresponds to very strong correlation. Nevertheless, the PL and GPL estimators behave in a qualitatively similar manner to the $\beta = 0.33$ case with small negative biases having a negligible contribution to the overall MSE. There is a slight indication, again, that the GPL estimators are slightly less variable than PL. For MCML estimation the parameter estimates of both α and β behave very badly for all choices of tuning constants when the number of Gibbs steps is 20,000 or less. Comparing GT1 to GT2 we see that by increasing the spacing while keeping the number of simulations fixed results can improve or deteriorate. When the spacing increases from 2 to 5 the estimates of α become less variable while those of β become more variable. When both spacing and number of simulation increase (compare GT1 and GT3) both parameter estimates become more stable, as was to be expected. Nevertheless, even for 20,000 simulations with a spacing of 20, MCMC estimates have approximately 40 – 60 times the standard error of the simple PL method. When the number of Gibbs steps is 200,000 we see that MCML behaves better, but parameter estimates are still more variable than either PL or GPL.

4 CONCLUSIONS

We have discussed estimation in the autologistic binary spatial model. Spatial models are naturally defined via conditional distributions, but for models that are defined conditionally there is an intractable normalizing constant in the joint likelihood that makes straightforward maximum likelihood estimation impractical. We have shown that at present there is no known simple way to approximate this normalizing constant. This leads to the need for alternative methods of estimation.

The Pseudo Likelihood (PL), Generalized Pseudo Likelihood (GPL), and Monte Carlo Maximum Likelihood (MCML) estimation techniques were discussed and applied to a binary spatial data set on cancer of the liver. We compared these methods of estimation in a simulation study and find that the more computationally intensive MCML method works well for weakly correlated spatial processes, but poorly for strongly correlated ones. The MCML method requires careful monitoring to get satisfactory results in the strong correlation setting while the PL and GPL methods are more “automatic”. In theory, the PL estimator has low efficiency relative to the ML estimator in the strong correlation setting. However, we see in Table 1 that PL is more efficient than the MCML method that attempts to approximate the MLE, at least for the number of MCML simulations we have used. Thus, large simulation experiments and/or careful monitoring of the Markov Chain is necessary to realize the benefits of the MCML approach.

For a given data set, however, the MCML approach gives standard errors of all parameter estimates together with the parameter estimates (and an estimated likelihood function). The PL and GPL methods do not automatically give standard errors, so nonparametric (model free) or parametric bootstrap methods (model based) that account for correlation are necessary. We have illustrated both methods of standard error estimation in the analysis of the cancer data set.

Although we have focused exclusively on binary data qualitatively similar results hold for other spatial data structures when the model is derived by conditioning. If we observe count data where each observation takes on one of K possible values, then arguing similarly to Section 2 the denominator term in Equation 2.2 has $K^{|D|}$ terms. Non-Gaussian continuous

variables lead to a $|D|$ variable integration. Thus, simple and direct Maximum Likelihood estimation is not feasible in these cases for large data sets either. This further demonstrates the need for alternative methods of estimation. The PL, GPL and MCML methods naturally apply to these other data structures.

REFERENCES

- Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society, Series B*, 23, 192-236.
- Besag, J. (1975). Statistical Analysis of Non-lattice Data, *The Statistician*, 24, 179-195.
- Comets, F. (1992). On Consistency of a Class of Estimators for Exponential Families of Markov Random Fields on the Lattice, *Annals of Statistics*, 20, 455-468.
- Cressie, N.A.C. (1993). *Statistics for Spatial Data*, Wiley, New York.
- Geman, S. and Graffigne, C. (1986). Markov Random Field Image Models and Their Applications to Computer Vision, in *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, Series B*, 54, 657-699.
- Gu, M. and Zhu, H. (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.
- 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.
- Hall, P. (1988). On Confidence Intervals for Spatial Parameters Estimated from Nonreplicated Data, *Biometrics*, 44, 271-277.
- Heagerty P.J. and Lumley, T. (2000). Window Subsampling of Estimating Functions With Application to Regression Models, *Journal of the American Statistical Association*, 95, 197-211.
- 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.
- Huffer, F.W. and Wu, H. (1998). Markov Chain Monte Carlo for Auto-logistic Regression Models with Applications to the Distribution of Plant Species, *Biometrics*, 54, 509-524.
- Künsch, H. (1989). The Jackknife and the Bootstrap for General Stationary Observations, *Annals of Statistics*, 17, 1217-1241.
- Liu, R. and Singh, K. (1992). Moving Blocks Jackknife and Bootstrap Capture Weak Dependence, in *Exploring the Limits of Bootstrap*, eds. Lepage, R. and Billard, L., John Wiley and Sons, N.Y.
- Lahiri, S.N. (1991). Second Order Optimality of Stationary Bootstrap, *Statistics and Probability Letters*, 11, 335-341.
- Moyeed, R.A. and Baddeley, A.J. (1991). Stochastic Approximation of the MLE for a Spatial Point Pattern, *Scandinavian Journal of Statistics*, 18, 39-50.
- Pickard, D. (1987). Inference for Discrete Markov Fields: the Simplest Nontrivial Case, *Journal of the American Statistical Association*, 82, 90-96.
- Politis, D. and Romano, J. (1994). Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions, *Annals of Statistics*, 22, 2031-2050.
- Riggan, W.B., Creason, J.P., Nelson, W.C., Manton, K.G., Woodbury, M.A., Stallard, E., Pellom, A.C., and Beaubier, J. (1987). *U.S. Cancer Mortality Rates and Trends, 1950-1979* (Vol. IV: Maps), U.S. Environmental Protection Agency, Washing, DC: U.S. Government Printing Office.
- Seymour, L. and Ji, C. (1996). Approximate Bayes Model Selection Procedures for Gibbs-Markov Random Fields, *Journal of Statistical Planning and Inference*, 51, 75-97.
- Sherman, M. and Carlstein, E. (1994). Nonparametric Estimation of the Moments of a General Statistic Computed from Spatial Data, *Journal of the American Statistical Association*, 89, 496-500
- Sherman, M. (1998). Efficiency and Robustness in Subsampling for Dependent Data, *Journal of Statistical Planning and Inference*, 75, 133-146.
- Snow, J. (1855). *On the Mode of Communication of Cholera (2nd ed.)*, Churchill, London.
- Younes, L. (1991). Maximum Likelihood Estimation for Gibbs Fields, *Spatial Statistics and*

Imaging (ed. A. Possolo), Lecture Notes, Monograph Series, Vol. 20, 403-426, Institute of Mathematical Statistics, Hayward, California.

nsim	burn-in	space	st.error($\hat{\alpha}$)	st.error($\hat{\beta}$)
2000	100	2	0.0360	0.0192
2000	100	5	0.0356	0.0193
2000	500	2	0.0361	0.0198
2000	500	5	0.0365	0.0200
5000	100	2	0.0365	0.0193
5000	100	5	0.0355	0.0190
5000	500	2	0.0366	0.0195
5000	500	5	0.0358	0.0193
8000	500	2	0.0360	0.0190
8000	500	5	0.0355	0.0187
8000	1000	2	0.0360	0.0190
8000	1000	5	0.0354	0.0186
10000	500	2	0.0362	0.0191
10000	500	5	0.0355	0.0186
10000	1000	2	0.0362	0.0191
10000	1000	5	0.0354	0.0185
8000	100	10	0.0357	0.0193
8000	100	20	0.0352	0.0190
8000	100	100	0.0359	0.0164
8000	500	10	0.0358	0.0194
8000	500	20	0.0352	0.0191
8000	500	100	0.0362	0.0165
10000	100	10	0.0356	0.0192
10000	100	20	0.0353	0.0187
10000	100	100	0.0358	0.0165
10000	500	10	0.0357	0.0193
10000	500	20	0.0353	0.0188
10000	500	100	0.0360	0.0165

Table 1: Parametric bootstrap SE's errors of Pseudo-Likelihood parameters for the cancer data set. The standard errors are given by $\text{st.error}(\hat{\alpha})$ and $\text{st.error}(\hat{\beta})$. The total number of Gibbs sampler steps is "nsim", "burn-in" denotes the number of Gibbs sampler steps until we accept the first simulated data set, and "space" is the number of Gibbs sampler steps taken between accepted simulated data sets.

scheme	Number		
	of terms	$\hat{\alpha}$	$\hat{\beta}$
g_{13}	1545	-0.3405	0.0962
g_9	1677	-0.3326	0.1000
g_5	1747	-0.3321	0.1011

Table 2: Parameter Estimates from Generalized Pseudo-Likelihood for the cancer data set. “scheme” denotes the size of sets, “Number of terms” is the number of terms in the Generalized Pseudo-Likelihood (product).

n	B	burn in	space	$\hat{\alpha}$	st.error($\hat{\alpha}$)	$\hat{\beta}$	st.error($\hat{\beta}$)
2000	951	100	2	-0.3032	0.0347	0.1175	0.0183
2000	381	100	5	-0.3040	0.0351	0.1173	0.0191
2000	751	500	2	-0.3032	0.0348	0.1171	0.0178
2000	301	500	5	-0.3047	0.0346	0.1166	0.0185
5000	2451	100	2	-0.3048	0.0336	0.1166	0.0182
5000	981	100	5	-0.3048	0.0357	0.1170	0.0189
5000	2251	500	2	-0.3049	0.0336	0.1164	0.0180
5000	901	500	5	-0.3051	0.0355	0.1168	0.0187
8000	3751	500	2	-0.3042	0.0336	0.1171	0.0182
8000	1501	500	5	-0.3038	0.0347	0.1176	0.0188
8000	3501	1000	2	-0.3044	0.0335	0.1169	0.0182
8000	1401	1000	5	-0.3038	0.0348	0.1176	0.0188
10000	4751	500	2	-0.3040	0.0340	0.1172	0.0184
10000	1901	500	5	-0.3034	0.0348	0.1178	0.0191
10000	4501	1000	2	-0.3041	0.0340	0.1171	0.0185
10000	1801	1000	5	-0.3034	0.0349	0.1178	0.0191

Table 3: Results from MCML estimation for the cancer data set: the estimates are $\hat{\alpha}$ and $\hat{\beta}$ and their standard errors are $\text{st.error}(\hat{\alpha})$ and $\text{st.error}(\hat{\beta})$. The total number of Gibbs sampler steps is n , B denotes the number of simulated data sets, the number of Gibbs sampler steps until we accept the first simulated data set is “burn in”, the number of Gibbs sampler steps taken between accepted simulated data sets is “space”.

$\alpha = 0.0, \beta = 0.33$				
method	$\hat{\alpha}$	st.error($\hat{\alpha}$)	$\hat{\beta}$	st.error($\hat{\beta}$)
ps	0.0005	0.0102	0.3285	0.0188
gps5	0.0001	0.0093	0.3285	0.0182
gps9	0.0001	0.0105	0.3263	0.0184
gps13	0.0004	0.0098	0.3285	0.0186
GT1	-0.0008	0.0203	0.3313	0.0336
GT2	-0.0013	0.0222	0.3301	0.0378
GT3	-0.0006	0.0164	0.3306	0.0283
GT4	-0.0006	0.0170	0.3307	0.0284
GT5	0.0026	0.0172	0.3279	0.0279
GT6	0.0029	0.0168	0.3285	0.0267
$\alpha = 0.0, \beta = 0.5$				
method	$\hat{\alpha}$	st.error($\hat{\alpha}$)	$\hat{\beta}$	st.error($\hat{\beta}$)
ps	0.0004	0.0437	0.4995	0.0298
gps5	0.0020	0.0409	0.4997	0.0294
gps9	0.0032	0.0405	0.4984	0.0291
gps13	0.0013	0.0417	0.4997	0.0295
GT1	-0.9638	8.3054	0.6466	2.5123
GT2	-0.1444	2.1908	0.5315	3.5444
GT3	-0.0278	1.7123	0.4657	0.6187
GT4	-0.0718	1.8505	0.3145	1.9751
GT5	-0.0015	0.0490	0.4902	0.0419
GT6	-0.0003	0.0489	0.4887	0.0474

Table 4: Simulation results: The number of simulated data sets is 100; the spacing between each is 500 Gibbs steps. “ps” is the pseudolikelihood estimate, “gps5” is generalized pseudolikelihood estimate with the number of sites in the set is $|g(i) = 5|$, “gps9” is generalized pseudolikelihood estimate with $|g(i) = 9|$, “gps13” is generalized pseudolikelihood estimate with $|g(i) = 13|$, “GT1” refers to MCML with $n = 5,000$ Gibbs steps, burn in of $d = 100$ and spacing of $\gamma = 2$; “GT2” refers to MCML with $n = 5,000$, $d = 100$ and $\gamma = 5$; “GT3” refers to MCML with $n = 20,000$, $d = 100$ and $\gamma = 10$; “GT4” refers to MCML with $n = 20,000$, $d = 100$ and $\gamma = 20$; “GT5” refers to MCML with $n = 200,000$, $d = 100$ and $\gamma = 20$; “GT6” refers to MCML with $n = 200,000$, $d = 100$ and $\gamma = 100$.

Figure Captions

Figure 1: Mortality map for cancer of the liver and gallbladder(including bile ducts) for white males during the decade from 1950-1959. Asterisks denote high rate counties and periods denote low rate counties.

Figure 2: The Pseudologlikelihood surface for the fitted model to the cancer data set.

Figure 3: The Monte Carlo Maximum loglikelihood surface for the fitted model to the cancer data set.

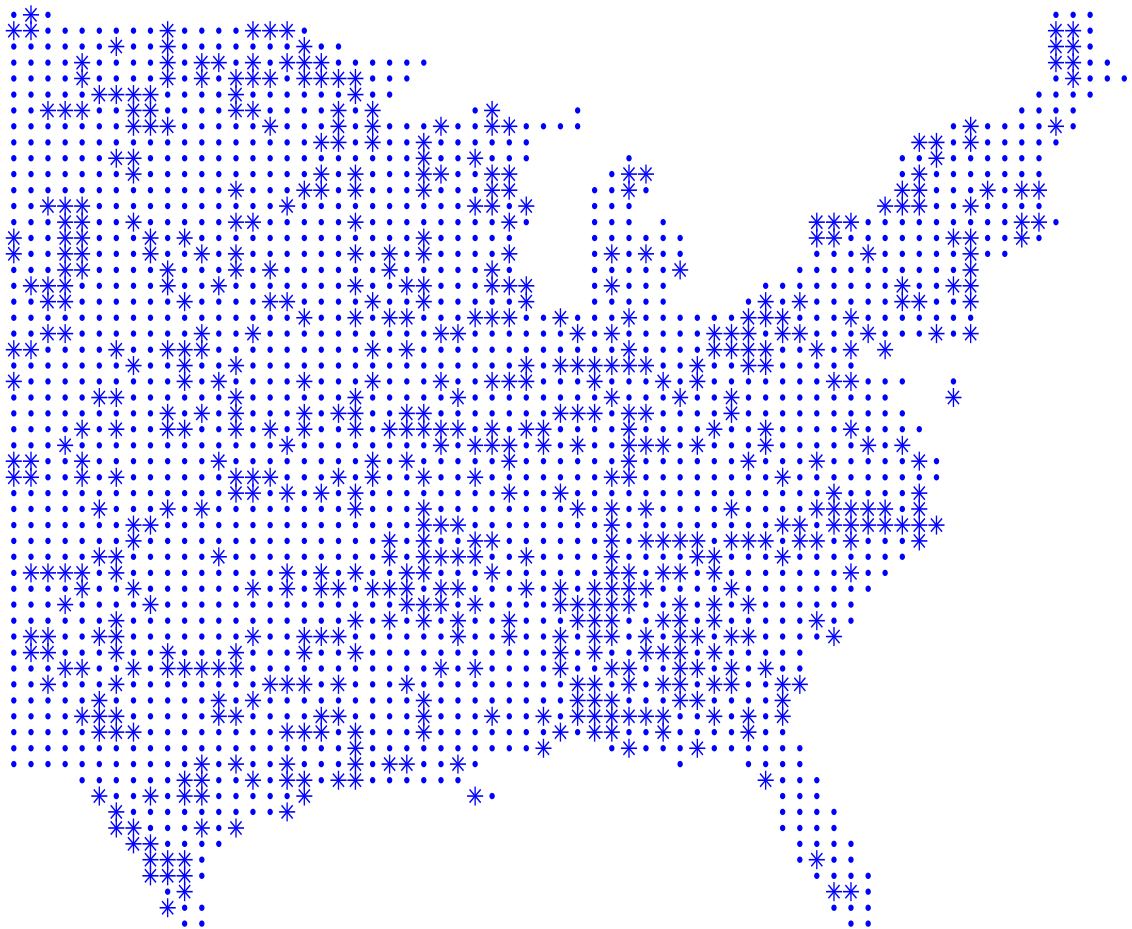


Figure 1:

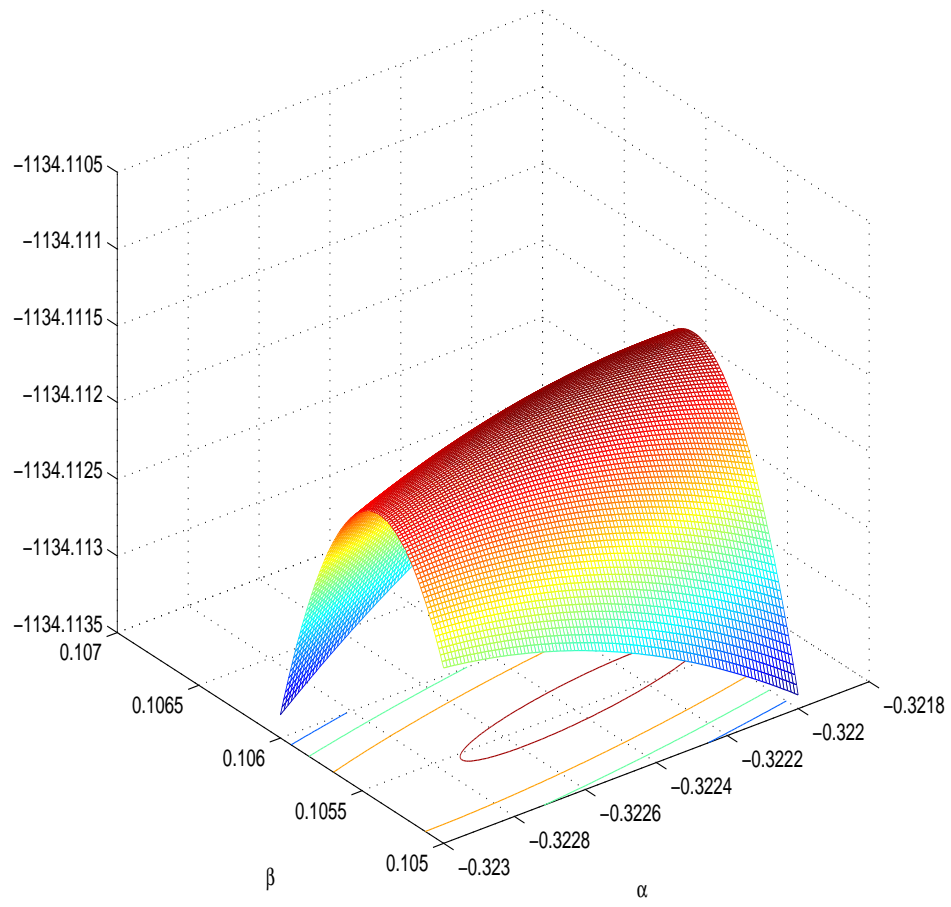


Figure 2:

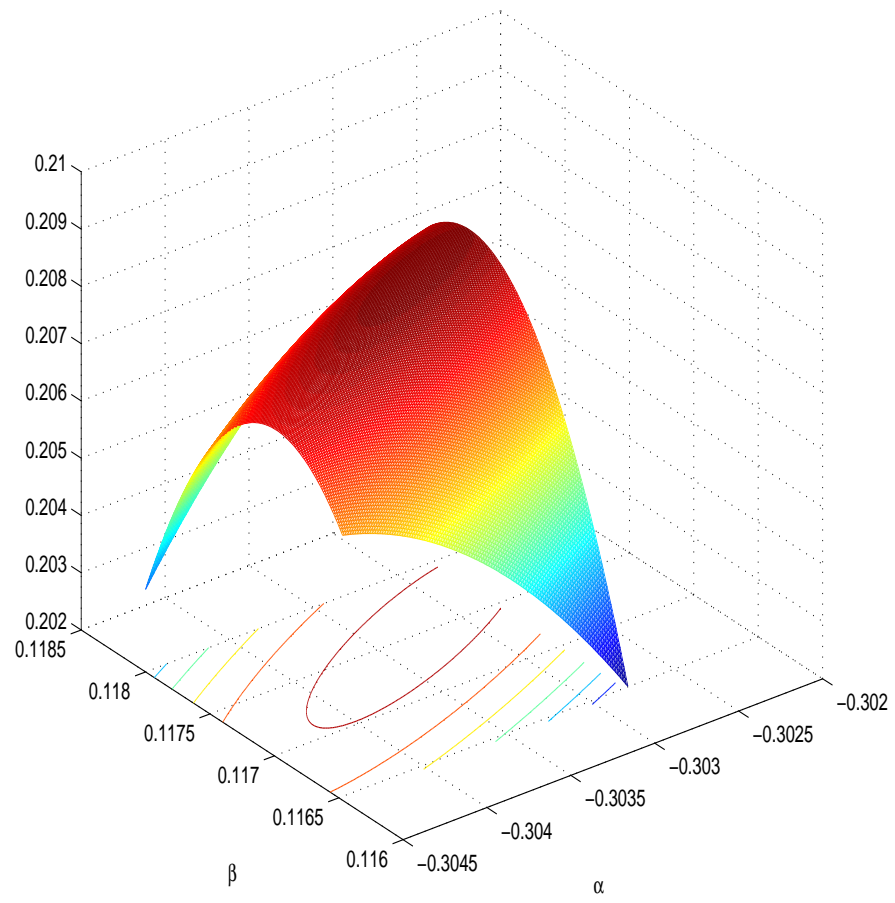


Figure 3: