A Double Metropolis-Hastings Sampler for Spatial Models with Intractable Normalizing Constants

Faming Liang [1]

Abstract

The problem of simulating from distributions with intractable normalizing constants has received much attention in the recent literature. In this paper, we propose an asymptotic algorithm, the so-called double Metropolis-Hastings (MH) sampler, for tickling this problem. Unlike other auxiliary variable algorithms, the double MH sampler removes the need of exact sampling, the auxiliary variables being generated using MH kernels, and thus can be applied to a wide range of problems for which exact sampling is not available. While for the problems for which exact sampling is available, it can typically produce the same accurate results as the exchange algorithm, but using much less CPU time. The new method is illustrated by various spatial models.

Keywords: Autologistic Model; Autonormal Model; Auxiliary Variable MCMC Algorithm; Exchange Algorithm; Metropolis-Hastings Algorithm.

1 Introduction

Spatial models, e.g., the autologistic model, the Potts model, and the autonormal model (Besag, 1974), have been used in modeling of many scientific problems. Examples include image analysis (Hurn *et al.*, 2003), disease mapping (Green and Richardson, 2002), genetic analysis (Francois *et al.*, 2006), among others. A major problem with the models is that the normalizing constant is intractable. The problem can be described as follows. Suppose we have a dataset $\boldsymbol{X}$ generated from a statistical model with the likelihood function

$$f(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} \exp\{-U(\boldsymbol{x}, \theta)\}, \quad \boldsymbol{x} \in \mathcal{X}, \quad \theta \in \Theta, \tag{1}$$

where θ is the parameter, and $Z(\theta)$ is the normalizing constant which depends on θ and is not available in closed form. Let $\pi(\theta)$ denote the prior density of θ. The posterior distribution of θ given $\boldsymbol{x}$ is then given by

$$\pi(\theta|\boldsymbol{x}) \propto \frac{1}{Z(\theta)} \exp\{-U(\boldsymbol{x}, \theta)\}\pi(\theta). \tag{2}$$

[1] Faming Liang is Associate Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA. Tel: +1-979-8458885; Email: fliang@stat.tamu.edu.

Since the closed form of $Z(\theta)$ is not available, inference from the model poses a great challenge on the current statistical methods.

The Metropolis-Hastings (MH) algorithm cannot be directly applied to simulate from $\pi(\theta|\boldsymbol{x})$, because the acceptance probability would involve the unknown ratio $Z(\theta)/Z(\theta')$, where θ' denotes the proposed value. To circumvent this difficulty, various approximation methods to the likelihood function or the normalizing constant function have been proposed for the models. The following are some examples. Besag (1974) proposed to approximate the likelihood function by a pseudo-likelihood function which is tractable. The method is easy to use, but it typically performs less well for the models for which neighboring dependence is strong. This method was further discussed and generalized by Dryden *et al.* (2002) and Huang and Ogata (2002). Geyer and Thompson (1992) proposed an importance sampling-based approach to approximate $Z(\theta)$. This approach was refined by Liang *et al.* (2007) by refining the choice of the trial density function using the stochastic approximation Monte Carlo algorithm. Liang (2007) proposed an alternative Monte Carlo approach to approximate $Z(\theta)$, where $Z(\theta)$ is viewed as a marginal distribution of the unnormalized distribution $g(\boldsymbol{x}, \theta) = \exp\{-U(\boldsymbol{x}, \theta)\}$ and is estimated by an adaptive kernel density estimator using Monte Carlo draws.

Recently, Møller *et al.* (2006) and Murray *et al.* (2006) proposed auxiliary variable MCMC algorithms for simulating from the distribution (2). These algorithms require exact sampling (Propp and Wilson, 1996) of $\boldsymbol{X}$. Unfortunately, exact sampling is very expensive or impossible for many statistical models whose normalizing constant is intractable.

In this paper, we propose a new asymptotic algorithm, the so-called double MH sampler, for simulating from the distributions with intractable normalizing constants. The double MH sampler removes the need of exact sampling, the auxiliary variables being generated using MH kernels, and thus can be applied to many statistical models for which exact sampling is not available. While for the models for which exact sampling is available, e.g., the autologistic model, it can produce almost the same accurate results as the exchange algorithm, but using much less CPU time.

The remainder of this paper is organized as follows. In Section 2, we give a brief review of the auxiliary variable MCMC algorithms. In Section 3, we describe the double MH sampler. In Section 4, we illustrate the double MH sampler with three spatial models, the autologistic model, the autonormal model, and the very-soft-core model. In Section 5, we conclude the paper with a brief discussion.

2 Auxiliary Variable MCMC Algorithms

In this section, we give a brief review of the auxiliary variable MCMC algorithms proposed by Møller *et al.* (2006) and Murray *et al.* (2006).

2.1 Møller *et al.*'s Algorithm

The key idea of Møller *et al.* (2006) is to extend the distribution $\pi(\theta|\boldsymbol{x})$ to include an auxiliary variable, $\boldsymbol{y}$, which shares the same state space as $\boldsymbol{x}$:

$$f(\theta, \boldsymbol{y}|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)\pi(\theta)f(\boldsymbol{y}|\theta, \boldsymbol{x}). \tag{3}$$

To simulate from (3) using the MH algorithm, the authors suggested the following proposal distribution

$$q(\theta', \boldsymbol{y}'|\theta, \boldsymbol{y}) = q(\theta'|\theta, \boldsymbol{y})q(\boldsymbol{y}'|\theta'), \tag{4}$$

which corresponds to the usual change in the parameter vector $\theta \to \theta'$, followed by exact sampling of $\boldsymbol{y}'$ from $q(\cdot|\theta')$. If $q(\boldsymbol{y}'|\theta')$ is set as $f(\boldsymbol{y}'|\theta)$, then the MH ratio can be written as

$$r(\theta, \boldsymbol{y}, \theta', \boldsymbol{y}'|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta')\pi(\theta')f(\boldsymbol{y}'|\theta', \boldsymbol{x})q(\theta|\theta', \boldsymbol{x})f(\boldsymbol{y}|\theta)}{f(\boldsymbol{x}|\theta)\pi(\theta)f(\boldsymbol{y}|\theta, \boldsymbol{x})q(\theta'|\theta, \boldsymbol{x})f(\boldsymbol{y}'|\theta')}, \tag{5}$$

where the unknown normalizing constant $Z(\theta)$ can be canceled. To ease computation, the authors further suggested to set the auxiliary distributions

$$f(\boldsymbol{y}|\theta, \boldsymbol{x}) = f(\boldsymbol{y}|\widehat{\theta}), \quad f(\boldsymbol{y}'|\theta', \boldsymbol{x}) = f(\boldsymbol{y}'|\widehat{\theta}), \tag{6}$$

where $\widehat{\theta}$ denotes an estimate of θ, for example, which can be obtained by maximizing a pseudo-likelihood function.

2.2 The Exchange Algorithm

Murray *et al.*'s algorithm is motivated by the parallel tempering algorithm (Geyer, 1991; Hukushima and Nemoto, 1996), and can be described as follows. Consider the augmented distribution

$$f(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_m, \theta|\boldsymbol{x}) = \pi(\theta)f(\boldsymbol{x}|\theta)\prod_{j=1}^{m} f(\boldsymbol{y}_j|\theta_j), \tag{7}$$

where θ_i's are instantiated and fixed, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are auxiliary variables with the same support as $\boldsymbol{x}$. Suppose that a change to θ is proposed with probability $q(\theta_i|\theta)$. To ensure that $\boldsymbol{y}_i = \boldsymbol{x}$, we swap the settings of $\boldsymbol{x}$ and $\boldsymbol{y}_i$. The resulting MH ratio for the change is

$$r(\theta, \theta_i, \boldsymbol{y}_i|\boldsymbol{x}) = \frac{\pi(\theta_i)f(\boldsymbol{x}|\theta_i)f(\boldsymbol{y}_i|\theta)\prod_{j\neq i} f(\boldsymbol{y}_j|\theta_j)q(\theta|\theta_i)}{\pi(\theta)f(\boldsymbol{x}|\theta)f(\boldsymbol{y}_i|\theta_i)\prod_{j\neq i} f(\boldsymbol{y}_j|\theta_j)q(\theta_i|\theta)} = \frac{\pi(\theta_i)f(\boldsymbol{x}|\theta_i)f(\boldsymbol{y}_i|\theta)q(\theta|\theta_i)}{\pi(\theta)f(\boldsymbol{x}|\theta)f(\boldsymbol{y}_i|\theta_i)q(\theta_i|\theta)}. \tag{8}$$

Based on the above arguments, the authors proposed the following algorithm:

The exchange algorithm

(a) Propose $\theta' \sim q(\theta'|\theta, \boldsymbol{x})$.

(b) Generate an auxiliary variable $\boldsymbol{y} \sim f(\boldsymbol{y}|\theta')$ using an exact sampler.

(c) Accept θ' with probability $\min\{1, r(\theta, \theta', \boldsymbol{y}|\boldsymbol{x})\}$, where

$$r(\theta, \theta', \boldsymbol{y}|\boldsymbol{x}) = \frac{\pi(\theta')f(\boldsymbol{x}|\theta')f(\boldsymbol{y}|\theta)q(\theta|\theta', \boldsymbol{x})}{\pi(\theta)f(\boldsymbol{x}|\theta)f(\boldsymbol{y}|\theta')q(\theta'|\theta, \boldsymbol{x})}. \tag{9}$$

Since a swapping change between $(\theta, \boldsymbol{x})$ and $(\theta', \boldsymbol{y})$ is involved, the algorithm is called the exchange algorithm by the authors. This algorithm represents an improvement over Møller *et al.*'s algorithm, as it removes the need to estimate the parameter before sampling begins. Murray *et al.* (2006) reported that the exchange algorithm tends to have a higher acceptance probability for the exact samples than Møller *et al.*'s algorithm.

3 A Double Metropolis-Hastings Sampler

Suppose that we are interested in simulating a sample $\boldsymbol{y}$ from $f(\boldsymbol{y}|\theta')$. If the sample is generated through m MH updates starting with the current state $\boldsymbol{x}$, the transition probability, $P_{\theta'}^{(m)}(\boldsymbol{y}|\boldsymbol{x})$, from $\boldsymbol{x}$ to $\boldsymbol{y}$ is then

$$P_{\theta'}^{(m)}(\boldsymbol{y}|\boldsymbol{x}) = K_{\theta'}(\boldsymbol{x} \to \boldsymbol{x}_1) \cdots K_{\theta'}(\boldsymbol{x}_{m-1} \to \boldsymbol{y}), \tag{10}$$

where $K(\cdot \to \cdot)$ is the MH transition kernel. Thus, we have

$$\begin{aligned} \frac{P_{\theta'}^{(m)}(\boldsymbol{x}|\boldsymbol{y})}{P_{\theta'}^{(m)}(\boldsymbol{y}|\boldsymbol{x})} &= \frac{K_{\theta'}(\boldsymbol{y} \to \boldsymbol{x}_{m-1}) \cdots K_{\theta'}(\boldsymbol{x}_1 \to \boldsymbol{x})}{K_{\theta'}(\boldsymbol{x} \to \boldsymbol{x}_1) \cdots K_{\theta'}(\boldsymbol{x}_{m-1} \to \boldsymbol{y})} \\ &= \frac{f(\boldsymbol{x}|\theta')}{f(\boldsymbol{y}|\theta')} \frac{f(\boldsymbol{y}|\theta')}{f(\boldsymbol{x}|\theta')} \frac{K_{\theta'}(\boldsymbol{y} \to \boldsymbol{x}_{m-1}) \cdots K_{\theta'}(\boldsymbol{x}_1 \to \boldsymbol{x})}{K_{\theta'}(\boldsymbol{x} \to \boldsymbol{x}_1) \cdots K_{\theta'}(\boldsymbol{x}_{m-1} \to \boldsymbol{y})} \\ &= \frac{f(\boldsymbol{x}|\theta')}{f(\boldsymbol{y}|\theta')}, \end{aligned} \tag{11}$$

where the last equality follows from the detailed balance equality $f(\boldsymbol{x}|\theta')K_{\theta'}(\boldsymbol{x} \to \boldsymbol{x}_1) \cdots K_{\theta'}(\boldsymbol{x}_{m-1} \to \boldsymbol{y}) = f(\boldsymbol{y}|\theta')K_{\theta'}(\boldsymbol{y} \to \boldsymbol{x}_{m-1}) \cdots K_{\theta'}(\boldsymbol{x}_1 \to \boldsymbol{x})$.

Now we return to the problem of simulating from the posterior distribution (2). By (11), the MH ratio (9) can be re-expressed as

$$r(\theta, \theta', \boldsymbol{y}|\boldsymbol{x}) = \frac{\pi(\theta')q(\theta|\theta', \boldsymbol{x})}{\pi(\theta)q(\theta'|\theta, \boldsymbol{x})} \frac{f(\boldsymbol{y}|\theta)P_{\theta'}^{(m)}(\boldsymbol{x}|\boldsymbol{y})}{f(\boldsymbol{x}|\theta)P_{\theta'}^{(m)}(\boldsymbol{y}|\boldsymbol{x})}. \tag{12}$$

It is easy to see that if we choose $q(\theta'|\theta, \boldsymbol{x})$ as a MH transition kernel which satisfies the detailed balance condition, then we have $\pi(\theta')q(\theta|\theta', \boldsymbol{x}) = \pi(\theta)q(\theta'|\theta, \boldsymbol{x})$, and the exchange update is reduced to a simple MH update for which $f(\boldsymbol{x}|\theta)$ works as the target distribution and $P_{\theta'}^{(m)}(\boldsymbol{y}|\boldsymbol{x})$ works as the proposal distribution. In summary, we have the following sampling scheme as a replacement for the exchange algorithm. Let t denotes the index of iterations, and let θ_t denote the current state of the Markov chain.

The double MH Sampler

(a) Simulate a new sample θ' from $\pi(\theta)$ using the MH algorithm starting with θ_t.

(b) Generate an auxiliary variable $\boldsymbol{y} \sim P_{\theta'}^{(m)}(\boldsymbol{y}|\boldsymbol{x})$, and accept it with probability $\min\{1, r(\theta_t, \theta', \boldsymbol{y}|\boldsymbol{x})\}$, where, by (11),

$$r(\theta_t, \theta', \boldsymbol{y}|\boldsymbol{x}) = \frac{f(\boldsymbol{y}|\theta_t)P_{\theta'}^{(m)}(\boldsymbol{x}|\boldsymbol{y})}{f(\boldsymbol{x}|\theta_t)P_{\theta'}^{(m)}(\boldsymbol{y}|\boldsymbol{x})} = \frac{f(\boldsymbol{y}|\theta_t)f(\boldsymbol{x}|\theta')}{f(\boldsymbol{x}|\theta_t)f(\boldsymbol{y}|\theta')}. \tag{13}$$

(c) Set $\theta_{t+1} = \theta'$ if the auxiliary variable is accepted in step (b), and set $\theta_{t+1} = \theta_t$ otherwise.

Since two types of MH updates are performed in step (b), one for drawing the auxiliary variable $\boldsymbol{y}$ and one for acceptance of θ', we call the algorithm the double MH sampler. Note that the MH update performed in step (a) is not essential, which can be incorporated into step (b) by changing (13) to (12). We also note that (13) holds regardless of the value of m. A remarkable feature of the algorithm is that it removes the need of exact sampling with a delicate use of the detailed balance condition. Therefore, the algorithm can be applied to a wide range of problems for which exact sampling is impossible or very expensive.

It is obvious that the samples will converge to the correct posterior distribution for a large value of m. In practice, to get good samples from the posterior distribution, m is not necessarily large. This can be justified as follows: Suppose that the current state θ_t is a sample from $\pi(\theta|\boldsymbol{x})$. If θ' is a good proposal, i.e., $\theta' \sim \pi(\theta|\boldsymbol{x})$, then we have $\boldsymbol{x} \sim f(\boldsymbol{x}|\theta')$. This further implies that $\boldsymbol{y} \sim f(\boldsymbol{y}|\theta')$ for any value of m, because $\boldsymbol{y}$ is generated through a sequence of MH updates which starts with $\boldsymbol{x}$ and admits $f(\cdot|\theta')$ as the invariant distribution. Hence, in this case, the transition $\theta_t \to \theta_{t+1}$ leaves the posterior distribution $\pi(\theta|\boldsymbol{x})$ invariant regardless the value of m. If θ' is a bad proposal, i.e., θ' is unlikely a sample from $\pi(\theta|\boldsymbol{x})$, then the ratio $f(\boldsymbol{x}|\theta')/f(\boldsymbol{y}|\theta')$ should be small, as $\boldsymbol{y}$ has moved some steps toward the equilibrium of $f(\cdot|\theta')$ than $\boldsymbol{x}$; and the ratio $f(\boldsymbol{y}|\theta_t)/f(\boldsymbol{x}|\theta_t)$ should also be small, as $\boldsymbol{y}$ has moved away from the equilibrium of $f(\cdot|\theta_t)$ than $\boldsymbol{x}$. This results in a very small MH ratio $r(\theta_t, \theta', \boldsymbol{y}|\boldsymbol{x})$. As a very likely consequence, the proposal is rejected and we set $\theta_{t+1} = \theta_t$. In all examples of this paper, the auxiliary variable $\boldsymbol{y}$ is generated through a single cycle of Gibbs iterations, and this translates to a value of m being equal to the dimension of $\boldsymbol{y}$. Note that the value of m used here is rather small, only one Gibbs update per component of $\boldsymbol{y}$. The key to the efficiency of the MH kernel $P_{\theta'}^{(m)}(\boldsymbol{y}|\boldsymbol{x})$ is of starting with $\boldsymbol{x}$, which also directly leads to the validity of (13) by the detailed balance condition.

Suppose that a sequence of samples $\theta_1, \ldots, \theta_n$ has been collected from a run of the double MH sampler. An approximate Bayesian estimator of θ can then be obtained by averaging over the samples,

$$\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n} \theta_i.$$

This estimator can also be named as an ensemble averaging estimator, as $\theta_1 \ldots, \theta_n$ are only approximately distributed as $\pi(\theta|\boldsymbol{x})$. As discussed by Haykin (1999, p.355), the ensemble averaging estimator has the same bias as, but a much smaller variance than the single sample estimator. This estimator can potentially be robustified by downweighting the samples for which the corresponding MH ratio $r(\theta_t, \theta', \boldsymbol{y}|\boldsymbol{x})$ was small or the corresponding proposal was rejected, with the reasons as explained above. Please refer to Haykin (1999) again for discussions on weight setting for ensemble averages. In this paper, we simply assign an equal weight to each sample, and call the resulting estimator $\bar{\theta}$ the approximate Bayesian estimator.

For an effective implementation of the double MH sampler, we need to consider two more issues. The first issue is on the choice of proposal distributions, namely, the proposal distribution used in step (a) for generating a new sample of θ', and the proposal distribution used in step (b) for generating an auxiliary variable of $\boldsymbol{y}$. As for conventional MCMC algorithms, these proposals should be adjusted such that they have a reasonable acceptance rate, e.g., a rate between 0.2 and 0.4 as suggested by Gelman *et al.* (1996). Since the double MH sampler undergoes two acceptance/rejection steps, a sample θ' is counted as acceptance only when acceptance is made in both steps (a) and (b). In the paper, we call the acceptance rate of θ' the acceptance rate of double MH moves. Note that when a Gibbs sampler is used for generating the auxiliary variables, the choice of the proposal is automatic.

The second issue is on diagnostic for the convergence of simulations. Since the double MH sampler belongs to the class of MCMC algorithms, its convergence can be monitored using the tools existing in the literature. In this paper, we adopted the the multiple run-based diagnostic method developed by Gelman and Rubin (1992).

4 Approximate Bayesian Analysis for Various Spatial Models

4.1 Spatial Autologistic Models

The autologistic model (Besag, 1974) has been widely used for spatial data analysis, see e.g., Preisler (1993), Wu and Huffer (1997), and Sherman *et al.* (2006). Let $\boldsymbol{x} = \{x_i : i \in D\}$ denote the observed binary data, where x_i is called a spin and D is the set of indices of the spins. Let $|D|$ denote the total number of spins in D, and let $n(i)$ denote the set of neighbors of spin i. The likelihood function of the model is

$$f(\boldsymbol{x}|\alpha,\beta) = \frac{1}{Z(\alpha,\beta)} \exp\left\{ \alpha \sum_{i\in D} x_i + \frac{\beta}{2} \sum_{i \in D} x_i \big(\sum_{j \in n(i)} x_j \big) \right\}, \quad (\alpha,\beta) \in \Theta, \tag{14}$$

where the parameter α determines the overall proportion of $x_i = +1$, the parameter β determines the intensity of interaction between x_i and its neighbors, and $Z(\alpha,\beta)$ is the intractable normalizing constant defined by

$$Z(\alpha,\beta) = \sum_{\text{for all possible } \boldsymbol{x}} \exp\left\{ \alpha \sum_{j\in D} x_j + \frac{\beta}{2} \sum_{i \in D} x_i \big(\sum_{j \in n(i)} x_j \big) \right\}.$$

An exact evaluation of $Z(\alpha,\beta)$ is impossible even for a moderate system.

To conduct a Bayesian analysis for the model, we assume a uniform prior on

$$(\alpha,\beta) \in \Theta = [-1,1] \times [0,1]$$

for all examples studied in §4.1. Then the double MH sampler can be applied to simulate from the posterior distribution $\pi(\alpha,\beta|\boldsymbol{x})$. In step (a), (α_t,β_t), the current state of the Markov chain, is updated by a single MH step with a random walk proposal $N_2((\alpha_t,\beta_t)', s^2 I_2)$, where s is the step size, and I_2 is the 2×2 identity matrix. In step (b), the auxiliary variable $\boldsymbol{y}$ is generated by a single cycle of Gibbs updates, and this translates to a value of $m = 2293$. Two or more cycles, which means a larger value of m, have also been tried for the examples, the results are similar. The acceptance rate of the double MH moves can be controlled by the choice of s. In this subsection, we set $s = 0.03$ for all examples.

4.1.1 U.S. Cancer Mortality Data

United States cancer mortality maps have been compiled by Riggan *et al* (1987) for investigating possible association of cancer with unusual demographic, environmental, industrial characteristics, or employment patterns. Figure 1(a) shows the mortality map of liver and gallbladder (including bile ducts) cancers for white males during the decade 1950-1959, which indicates some apparent geographic clustering. Refer to Sherman *et al.* (2006) for more descriptions of the data. Following Sherman *et al.* (2006), we modeled the data by a spatial autologistic model. The total number of spins is $|D| = 2293$. A free boundary condition is assumed for the model, under which the boundary points have less neighboring points than the interior points. The assumption is natural to this example, as the lattice has an irregular shape.

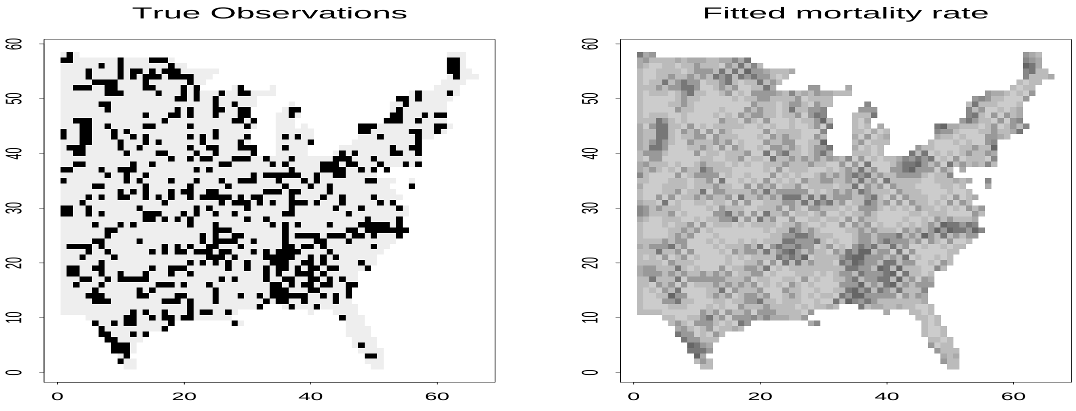


Figure 1: US cancer mortality data. (a) The mortality map of liver and gallbladder cancers (including bile ducts) for white males during the decade 1950-1959. Black squares denote counties of high cancer mortality rate, and white squares denote counties of low cancer mortality rate. (b) Fitted cancer mortality rates by the autologistic model with the parameters being replaced by its approximate Bayesian estimates. The cancer mortality rate of each county is represented by the gray level of the corresponding square.

The double MH sampler was first applied to this example. The sampler started with the initial value $(\alpha_0, \beta_0) = (0, 0)$ and was run 5 times independently. Each run consisted of 10500 iterations. The CPU time cost by each run was 4.2s on a 2.8GHz computer (all computations reported in this paper were done in the same computer). The overall acceptance rate of the double MH moves was about 0.23. Figure 2 provides two diagnostic plots for the convergence of the runs, where the statistic Gelman-Rubin $\hat{R}$ (Gelman and Rubin, 1992) was plotted against iterations. The simulations are usually considered as converged, when the statistic Gelman-Rubin $\hat{R}$ falls below the horizontal line 1.1. Figure 2 indicates that for this example, the simulations converged very fast, usually within two hundreds of iterations. Based on this diagnostic, we discarded the first 500 iterations of each run for the burn-in process, and collected 2000 samples from the remaining iterations at equally spaced time points. Averaging over the estimates obtained from respective runs, we got the following estimate: $(\widehat{\alpha}, \widehat{\beta}) = (-0.3028, 0.1228)$ with the standard error $(8.2 \times 10^{-4}, 2.7 \times 10^{-4})$.

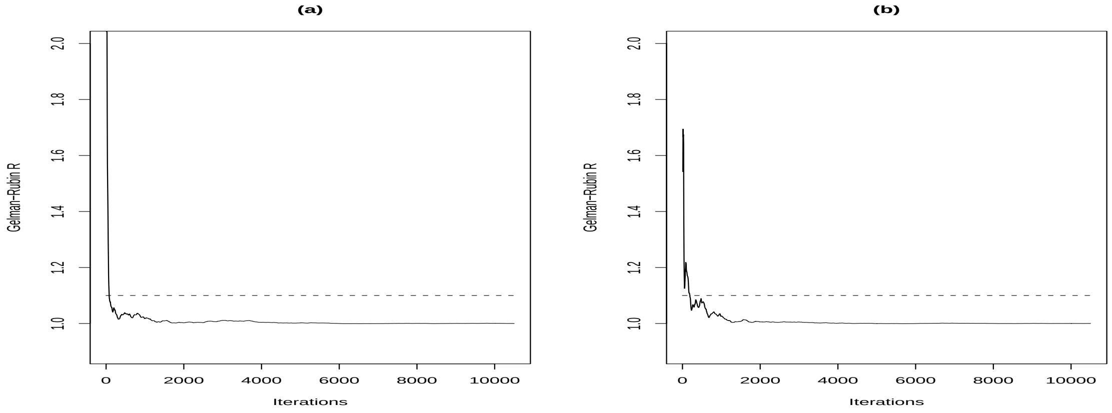


Figure 2: Convergence diagnostic of the double MH sampler for the U.S. Cancer Mortality example: (a) Gelman-Rubin diagnostic based on the samples of α generated in 5 runs; (b) Gelman-Rubin diagnostic based on the samples of β generated in 5 runs.

For comparison, the exchange algorithm was also applied to this example. It was run as the double MH sampler except that the auxiliary variable $\boldsymbol{y}$ was generated using an exact sampler. Following Murray *et al.* (2006), we adopted the summary state algorithm (Childs *et al.*, 2001) as our exact sampler, which is suitable for high dimensional binary spaces. The algorithm was also run 5 times, and each run consisted of 10500 iterations. The CPU time cost by each run was 111.5s, about 27 times longer than that cost by the double MH sampler. The overall acceptance rate of the exact auxiliary variables was 0.2. Averaging over the estimates obtained from respective runs, we got the estimate $(\widehat{\alpha}, \widehat{\beta}) = (-0.3030, 0.1219)$ with the standard error $(1.1 \times 10^{-3}, 6.0 \times 10^{-4})$.

It is easy to see that the double MH sampler and the exchange algorithm produced almost identical estimates for this example. These estimates are also very close to the estimate $(-0.3008, 0.1231)$ obtained by Liang (2007) using contour Monte Carlo, and the estimate $(-0.2999, 0.1234)$ obtained by Liang *et al.* (2007) using stochastic approximation Monte Carlo. We note that both the contour Monte Carlo and stochastic approximation Monte Carlo algorithms try to first approximate the unknown normalizing constant function, and then estimate the parameters based on the approximated normalizing constant function. As reported by the authors, both the algorithms take hours of CPU time to approximate the normalizing constant function. This data has also been analyzed by Sherman *et al.* (2006) using the Monte Carlo maximum likelihood algorithm (Geyer and Thompson, 2002), resulting in a similar estimate of $(-0.304, 0.117)$.

Later, both the double MH sampler and the exchange algorithm were re-run with the same parameter setting as specified above except for the number of iterations being lengthened to 100500. In each run, 10000 samples were collected from the last 100000 iterations at equally spaced time points. The empirical distributions of the samples were studied in Figure 3. The plots indicate that the samples generated by the double MH sampler are almost identically distributed as those generated by the exchange algorithm.

In summary, for this example the double MH sampler produced almost identical results with the exchange algorithm while using much less CPU time. This advantage can be seen clearer in §4.1.2, where for some cases the exact sampler is impossible while the double MH sampler still works well.

4.1.2 Simulation Studies

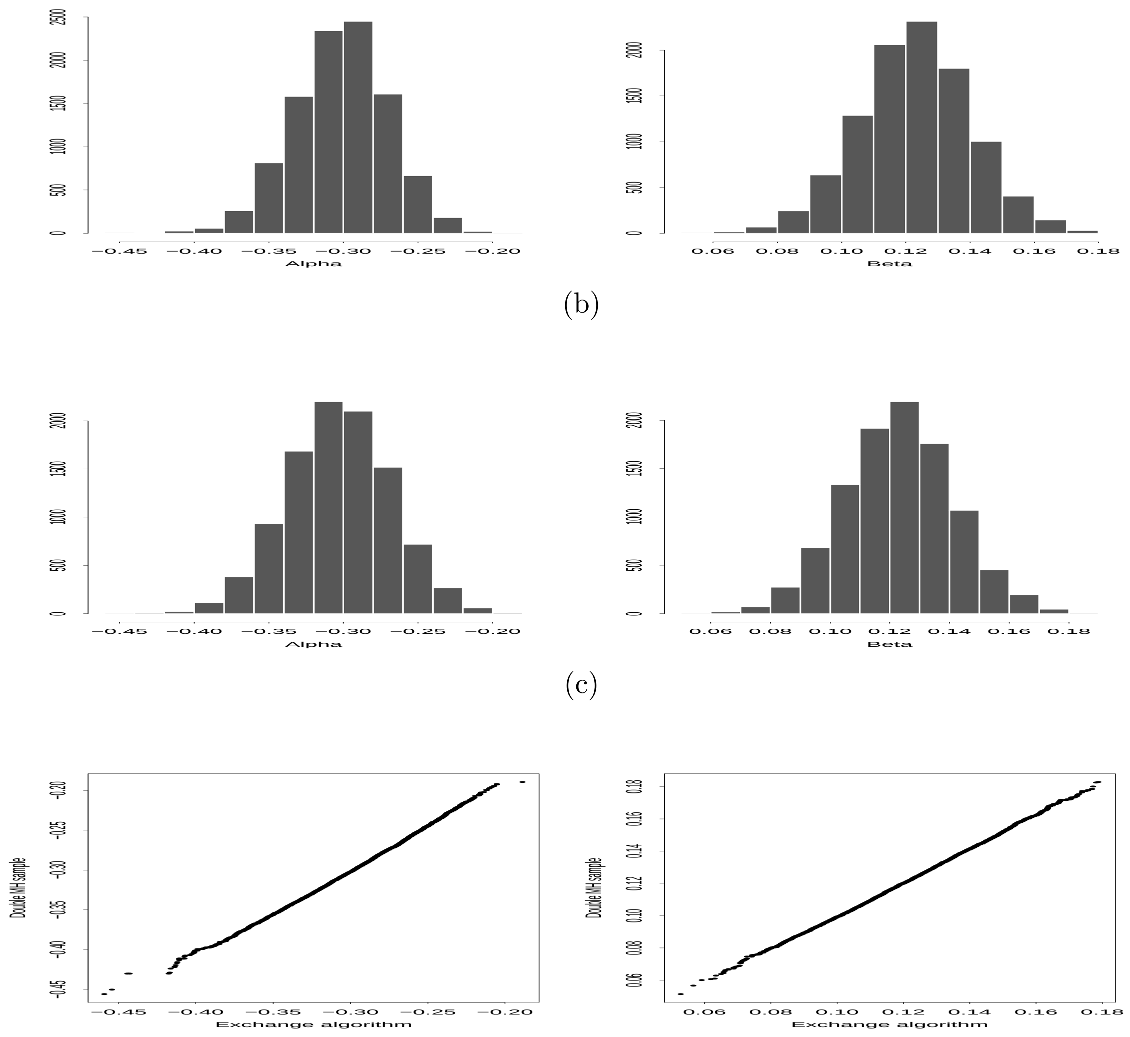
To assess the general accuracy of the estimates produced by the double MH sampler, we simulated 50 independent samples for the U.S. cancer mortality data under each setting of (α, β) given in Table 1. Since the lattice is irregular, the free boundary condition was again assumed in the simulations. We then re-estimated the parameters using the double MH sampler and the exchange algorithm. Both algorithms were run as in §4.1.1. The computational results were summarized in Table 1. For a thorough comparison, we also included in Table 1 the maximum pseudo-likelihood estimators (MPLE) of the parameters (Besag, 1974), which were obtained by maximizing the pseudo-likelihood function

$$\tilde{L}(\alpha, \beta|\boldsymbol{x}) = \prod_{i \in D} \left[\frac{e^{\alpha x_i + \beta \sum_{j \in n(i)} x_i x_j}}{e^{\alpha + \beta \sum_{j \in n(i)} x_j} + e^{-\alpha - \beta \sum_{j \in n(i)} \sum x_j}} \right]$$

using the downhill simplex method (Press *et al.*, 1992). The advantage of this method is that it does not require the gradient information of the objective function, and can thus be easily applied to the constraint optimization problems.

Table 1 indicates that the double MH sampler can produce almost the same accurate results as does the exchange algorithm, and more accurate results than does the MPLE especially when α and β are large. It is remarkable that the CPU time cost by the double MH sampler is independent of the values of (α, β). Whilst as β increases, the CPU time cost by the exchange algorithm increases exponentially. Childs *et al.* (2006) studied the behavior of the exact sampler for the Ising model, a simplified autologistic model. For the Ising model, they fitted an exponential law for the convergence time, and reported that the exact sampler may diverge at a value of β lower than the critical value (≈ 0.44). Childs *et al.*'s finding is consistent with our results. It takes an extremely long CPU time

(a)



(b)

(c)

Figure 3: Panel (a) shows the histograms of the samples generated by the exchange algorithm. Panel (b) shows the histograms of the samples generated by the double MH sampler. Panel (c) shows the Q-Q plots for the samples generated by these two algorithms, the left plot is for the α samples, and the right plot is for the β samples.

Table 1: Computational results for the simulated U.S. cancer mortality data. The numbers in the parentheses denote the standard error of the estimates. Notes: —, not available; [a] the CPU time cost by a single run of the double MH sampler; [b] the CPU time cost by a single run of the exchange algorithm; [e] the data were simulated using the exact sampler; [g] the data were simulated using the Gibbs sampler, starting with a random configuration and then iterating for 100000 Gibbs cycles.

	Double MH sampler			Exchange algorithm			MPLE	
(α, β)	$\widehat{\alpha}$	$\widehat{\beta}$	CPU^a(s)	$\widehat{\alpha}$	$\widehat{\beta}$	CPU^b(s)	$\widehat{\alpha}$	$\widehat{\beta}$
(0,0.1)^e	−.0038	.1010	4.2	−.0038	.1002	103	-.0035	.1016
	(.0024)	(.0018)		(.0024)	(.0018)		(.0024)	(.0019)
(0,0.2)^e	−.0026	.2018	4.2	−.0025	.2007	251	-.0024	.2025
	(.0021)	(.0019)		(.0020)	(.0019)		(.0022)	(.0022)
(0,0.3)^e	−.0018	.2994	4.2	−.0014	.2971	821	-.0019	.2981
	(.0014)	(.0018)		(.0014)	(.0018)		(.0016)	(.0022)
(0,0.4)^e	.0013	.4023	4.2	−.0007	.3980	7938	.0020	.4013
	(.0009)	(.0015)		(.0004)	(.0012)		(.0012)	(.0020)
(0.1,0.1)^e	.1025	.0993	4.2	.1030	.0986	110	.1023	.0999
	(.0025)	(.0022)		(.0025)	(.0022)		(.0025)	(.0023)
(0.3,0.3)^e	.2944	.3032	4.2	.3012	.3008	321	.2904	.3041
	(.0098)	(.0043)		(.0098)	(.0043)		(.0102)	(.0046)
(0.5,0.5)^g	.5040	.5060	4.2	—	—	—	.5610	.4847
	(.0227)	(.0085)		—	—		(.0393)	(.0123)

for the exact sampler to generate a sample under the settings $(0, 0.4)$ and $(0.5, 0.5)$. We note that due to the effect of α, it usually takes a longer CPU time for the exact sampler to generate a sampler under the setting $(0, \beta)$ than under the setting (α, β); and that when α and β are large, the accuracy of the estimates tend to be reduced by their correlation.

4.2 Autonormal Models

Consider a second-order zero-mean Gaussian Markov random field $\boldsymbol{X} = (X_{ij})$ defined on an $M \times N$ lattice, whose conditional density function is given by

$$\begin{aligned} f(x_{ij}|\boldsymbol{\beta}, \sigma^2, x_{uv}; (u,v) \neq (i,j)) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ &- \frac{1}{2\sigma^2}(x_{ij} - \beta_h \sum_{(u,v)\in n_h(i,j)} x_{uv} \\ &- \beta_v \sum_{(u,v)\in n_v(i,j)} x_{uv} - \beta_d \sum_{(u,v)\in n_d(i,j)} x_{uv})^2 \Big\}, \end{aligned} \tag{15}$$

where $\boldsymbol{\beta} = (\beta_h, \beta_v, \beta_d)$ and σ^2 are parameters, $n_h(i,j) = \{(i, j-1), (i, j+1)\}$, $n_v(i,j) = \{(i-1, j), (i+1, j)\}$ and $n_d(i,j) = \{(i-1, j-1), (i-1, j+1), (i+1, j-1), (i+1, j+1)\}$ are neighbors of (i,j). This model is stationary when $|\beta_h| + |\beta_v| + 2|\beta_d| < 0.5$ (Balram and Moura, 1993). The joint likelihood function of this model can be written as

$$f(\boldsymbol{x}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-MN/2}|B|^{1/2} \exp\{-\frac{1}{2\sigma^2}\boldsymbol{x}' B \boldsymbol{x}\},$$

where B is an $(MN \times MN)$-dimensional matrix, e.g., a 2500×2500 matrix corresponding to a small lattice of size 50×50, and $|B|$ is intractable except for some special cases (Besag and Moran, 1975).

To conduct a Bayesian analysis for the model, we assume the following priors:

$$\pi(\boldsymbol{\beta}) \propto I(|\beta_h| + |\beta_v| + 2|\beta_d| < 0.5), \qquad \pi(\sigma^2) \propto \frac{1}{\sigma^2}, \tag{16}$$

which $I(\cdot)$ is the indicator function. Under the free boundary condition for which the boundary pixels have fewer neighbors, we have the following posterior distribution

$$\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{x}) \propto (\sigma^2)^{-\frac{MN}{2}-1}|B|^{1/2} \exp\Big\{-\frac{MN}{2\sigma^2}\Big(S_x - 2\beta_h X_h - 2\beta_v X_v - 2\beta_d X_d\Big)\Big\} I(|\beta_h| + |\beta_v| + 2|\beta_d| < 0.5), \tag{17}$$

where

$$S_x = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} x_{ij}^2, \quad X_h = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N-1} x_{ij}x_{i,j+1},$$

$$X_v = \frac{1}{MN}\sum_{i=1}^{M-1}\sum_{j=1}^{N} x_{ij}x_{i+1,j}, \quad X_d = \frac{1}{MN}\Big(\sum_{i=1}^{M-1}\sum_{j=1}^{N-1} x_{ij}x_{i+1,j+1} + \sum_{i=1}^{M-1}\sum_{j=2}^{N} x_{ij}x_{i+1,j-1}\Big).$$

Although σ^2 can be integrated out from the posterior, we do not suggest to do so. Working on the joint posterior will ease the generation of auxiliary variables for the double MH sampler. To ease implementation of sampling from the prior distribution, we reparametrize σ^2 by $\tau = \log(\sigma^2)$, and then we have

$$\pi(\boldsymbol{\beta}, \tau) \propto I(|\beta_h| + |\beta_v| + 2|\beta_d| < 0.5).$$

The double MH sampler can be applied to simulate from the posterior distribution of $\boldsymbol{\beta}$ and τ. In step (a), $(\boldsymbol{\beta}_t, \tau_t)$ is updated by a single MH step with a random walk proposal $N((\boldsymbol{\beta}_t, \tau_t)', s^2 I_4)$. In this subsection, we set $s = 0.02$ unless otherwise stated. In step (b), the auxiliary variable $\boldsymbol{y}$ is generated by a single cycle of Gibbs updates:

$$y_{ij}|\boldsymbol{y}_{(u,v)\in n(i,j)} \sim N\left(\beta_h \sum_{(u,v)\in n_h(i,j)} y_{uv} + \beta_v \sum_{(u,v)\in n_v(i,j)} y_{uv} + \beta_d \sum_{(u,v)\in n_d(i,j)} y_{uv}, e^{\tau_t}\right),$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, starting with $\boldsymbol{y} = \boldsymbol{x}$.

The exchange algorithm is not applicable to the autonormal model, as exact sampling is not available for it. However, under the free boundary condition, the log-likelihood function of the model admits the following analytic form (Balram and Moura, 1993):

$$\begin{aligned} l(\boldsymbol{X}|\boldsymbol{\beta}, \sigma^2) &= \text{Constant} - \frac{MN}{2}\log(\sigma^2) - \frac{MN}{2\sigma^2}\Big(S_x - 2\beta_h X_h - 2\beta_v X_v - 2\beta_d X_d\Big) \\ &\quad + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} \log\left(1 - 2\beta_v \cos\frac{i\pi}{M+1} - 2\beta_h \cos\frac{j\pi}{N+1} - 4\beta_d \cos\frac{i\pi}{M+1}\cos\frac{j\pi}{N+1}\right), \end{aligned} \tag{18}$$

where S_x, X_h, X_v and X_d are as defined in (17). The Bayesian inference for the model is then standard, with the priors as specified in (16). In this paper, we call the Bayesian analysis based on this analytic likelihood function the true Bayesian analysis, and call the resulting estimator the true Bayesian estimator.

For a thorough comparison, we also considered MPLEs for this model, which are to find the parameter values that maximize the pseudo-likelihood function:

$$\widetilde{L}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{x}) = \prod_{i=1}^{M}\prod_{j=1}^{N} f(x_{ij}|\boldsymbol{\beta}, \sigma^2, x_{uv}; (u,v) \neq (i,j)).$$

The maximization can be accomplished using the downhill simplex method (Press *et al.*, 1992).

4.2.1 Wheat Yield Data

We first work on the wheat yield data collected on a 20 × 25 rectangular lattice (Table 6.1, Andrews and Herzberg, 1985). The data was shown in Figure 4(a), which indicates positive correlation between neighboring observations. This data has been analyzed by a number of authors, e.g., Besag (1974), Huang and Ogata (1999) and Gu and Zhu (2001). Following the previous authors, we subtracted the mean from the data and then fitted them by the autonormal model. In our analysis, the free boundary condition is assumed. This is natural, as for the real data the lattice is often irregular.

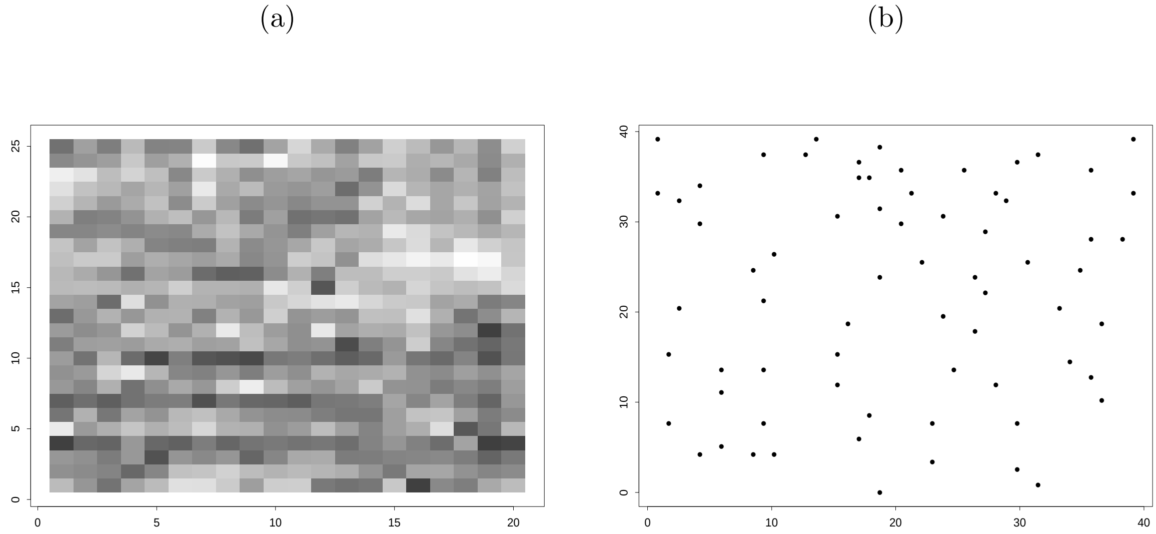


Figure 4: (a) Image of the wheat yield data: black squares denote high yield ares, and white squares denote low yield areas. (b) Locations of 69 Spanish town in an area of 40 × 40 square miles.

The double MH sampler is applied to this example. The sampler was run 5 times independently. Each run started with the point (0,0,0,0) and consisted of 50500 iterations, for which the first 500 iterations were discarded for the burn-in process and 10000 samples were collected from the remaining iterations at equally spaced time points. The CPU time cost by each run was 5.8s. The overall acceptance rate of the double MH moves was about 0.23. The results were summarized in Table 2. To assess the quality of the estimators, we also included in the table the mean squared error of the fitted values, which is defined as

$$\text{FMSE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{ij} - \hat{x}_{ij}),$$

and $\hat{x}_{ij}$ denotes the fitted value of x_{ij}. FMSE provides a square measure for the difference between the fitted and true observations. For the Bayesian method, $\hat{x}_{ij}$ can be calculated by

$$\hat{x}_{ij} = \frac{1}{n}\sum_{t=1}^{n}\left(\beta_h^{(t)} \sum_{(u,v)\in n_h(i,j)} x_{uv} + \beta_v^{(t)} \sum_{(u,v)\in n_v(i,j)} x_{uv} + \beta_d^{(t)} \sum_{(u,v)\in n_d(i,j)} x_{uv}\right),$$

given the MCMC samples $(\boldsymbol{\beta}^{(t)}, \tau^{(t)})$, $t = 1, \ldots, n$.

Table 2: Computational results for the wheat yield data. The numbers in the parentheses denote the standard error of the estimates.

Algorithm	β_h	β_v	β_d	σ^2	FMSE
True Bayes	0.102(4e-4)	0.355(3e-4)	0.006(2e-4)	0.123(2e-4)	0.123(0.0)
DMH	0.099(6e-4)	0.351(5e-4)	0.006(3e-4)	0.126(3e-4)	0.123(0.0)
MPLE	0.140	0.340	-0.010	0.122	0.122

For comparison, the true Bayesian and MPLE methods were also applied to this example. In the true Bayesian analysis, the posterior was simulated using the MH algorithm 5 times. Each run also consisted of 50500 iterations, with he first 500 iterations being discarded for the burn-in process and 10000 samples being collected from the remaining iterations at equally spaced time points. The proposal adopted here was a random walk proposal with the variance-covariance matrix $0.02^2 I_4$. The overall acceptance rate of the proposals was about 0.22. The numerical results were also summarized in Table 2. The comparison indicates that the estimates produced by DMH are much closer to the true Bayesian estimates than are MPLEs, while the three methods all produced about the same FMSEs.

4.2.2 Simulation Studies

To assess the general accuracy of the double MH estimator for the autonormal models, we simulated 50 independent samples of size 100 × 100 under each setting of $(\boldsymbol{\beta}, \sigma^2)$ given in Table 3. Without loss of generality, we set $\sigma^2 = 1.0$. The simulations were done using the Gibbs sampler, starting with a random configuration with each entry being drawn independently from $N(0,1)$ and then iterating for 50000 Gibbs cycles. The free boundary condition was again assumed in the simulations. The parameters were re-estimated using the double MH sampler, the true Bayesian method, and the MPLE method. These methods were run as in §4.2.1 except for the choice of s. Here we set $s = 0.005$ and this resulted in an acceptance rate of 0.35 for the double MH sampler and an acceptance rate of 0.37 for the true Bayesian method. The numerical results were summarized in Table 3. Instead of FMSEs, we considered for this example the predictive log-score (Hoeting *et al.*, 1999), which is defined as

$$-\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\log\left\{\frac{1}{n}\sum_{t=1}^{n}P\Big(z_{ij}|\boldsymbol{\beta}^{(t)},\tau^{(t)},z_{uv};(u,v)\neq(i,j)\Big)\right\},$$

Table 3: Computational results for the simulated auto-normal data. All entries of the table have been scaled by a factor of 1000. The true parameter values $(\beta_h, \beta_v, \beta_d, \sigma^2)$ are (0.1,0.1,0.0,1), (0.15,0.1,0.1,1.0) and (0.2,0.15,0.05,1.0) for the settings I, II and III, respectively. Let θ be the true value of a parameter, and let $\hat{\theta}^{(i)}$ denote its estimate obtained from dataset i, then "bias"=$\sum_{i=1}^{50}(\hat{\theta}^{(i)} - \theta)/50$, "se" is the standard error of $\sum_{i=1}^{50} \hat{\theta}^{(i)}/50$, and "rmse"=$\sqrt{\sum_{i=1}^{50}(\hat{\theta}^{(i)} - \theta)^2/50}$.

Algorithm		True Bayesian method			DMH			MPLE		
Setting		I	II	III	I	II	III	I	II	III
β_h	bias	0.34	-0.75	3.04	-0.29	-1.42	1.60	0.22	-1.96	2.68
	se	1.55	1.22	1.12	1.52	1.22	1.14	1.63	1.46	1.25
	rmse	10.83	8.56	8.44	10.66	8.67	8.14	11.39	10.40	9.18
β_v	bias	-1.73	0.92	0.35	-2.28	0.73	-0.32	-2.65	1.77	-1.39
	se	1.25	1.26	1.41	1.24	1.25	1.41	1.31	1.85	1.67
	rmse	8.95	8.90	9.91	8.98	8.80	9.89	9.52	13.06	11.75
β_d	bias	-0.44	0.24	-1.15	-0.23	0.57	0.16	-0.39	0.36	-0.63
	se	1.08	0.88	1.04	1.08	0.84	2.30	1.11	0.92	1.17
	rmse	7.60	6.15	7.36	7.57	5.94	7.57	7.79	6.48	8.23
σ^2	bias	-4.08	-13.02	-14.01	0.36	-0.48	0.16	0.05	-1.68	2.18
	se	2.02	2.46	2.27	2.05	2.43	2.30	2.05	2.93	2.90
	rmse	14.69	21.57	21.21	14.36	17.02	16.07	14.36	20.55	20.43
Score	mean	1418.7	1418.7	1419.1	1418.7	1418.7	1419.1	1418.8	1418.9	1419.2
	se	0.96	0.99	1.08	0.96	0.97	1.07	0.96	0.96	1.07

given the MCMC samples $(\boldsymbol{\beta}^{(t)}, \tau^{(t)})$, $t = 1, \ldots, n$. Here $\boldsymbol{z}$ denotes a sample simulated independently of $\boldsymbol{x}$ but under the same parameter setting. As argued by Hoeting *et al.* (1999), the predictive score is a combined measure of the predictive bias and the lack of calibration. The smaller the score, the better the predictive performance.

Table 3 indicates that the double MH sampler outperforms the MPLE method for this example; the Double MH estimates consistently have smaller biases, standard errors, root mean square errors, and predictive log-scores than the MPLEs. It is remarkable that for the parameters β_h, β_v and β_d, the DMH estimates are quite comparable with the true Bayesian estimates; and for σ^2, the DMH estimate is even better than the true Bayesian estimate.

4.3 Pairwise Interacted Spatial Point Process

The spatial point process is described by the coordinates of points $\boldsymbol{x} = \{x_i \in A : i = 1, \ldots, n\}$ in a planar region A. If the points are pairwise interacted, the joint density can be written as

$$f(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} \exp\left\{ -\sum_{i=1}^{n} \sum_{j>i} \phi(\|x_i - x_j\|, \theta) \right\}, \quad \theta > 0, \tag{19}$$

where $\phi(\cdot)$ is called the pairwise potential function. If we define

$$\phi(t, \theta) = -\log\{1 - \exp(-\rho t^2/\theta)\},$$

then the model (19) is called the very-soft-core model, where $\rho = n/|A|$ and $|A|$ is the area of region A. The normalizing constant of the model is

$$Z(\theta) = \int_A \cdots \int_A \exp\{-\sum_{i=1}^{n} \sum_{j>i} \phi(\|x_i - x_j\|, \theta)\} dx_1 \cdots dx_n,$$

which is intractable. To estimate the parameter θ, various approximations to the function $Z(\theta)$ have been proposed in the literature, see Huang and Ogata (1999) for an overview.

To conduct a Bayesian analysis for the model, we let θ be subject to the following prior:

$$\pi(\theta) \propto 1/\theta,$$

and then reparametrize it by $\tau = \log(\theta)$. The double MH sampler can then be applied to simulate from the posterior density of τ. In step (a), τ_t, the current state of the Markov chain, is updated by a MH step with a random walk proposal $N(\tau_t, s^2)$, where we set $s = 1.5$ for the Spanish town example studied below. In step (b), the auxiliary variable $\boldsymbol{y}$ is simulated through a single cycle of Metropolis-within-Gibbs moves (Müller, 1993):

$$y_i \sim f(y_i|\boldsymbol{y}_{[-i]}) \propto \exp\{-\sum_{j \neq i} \phi(\|y_j - y_i\|, e^{\tau})\}, \quad i = 1, \ldots, n, \tag{20}$$

where $\boldsymbol{y}$ is initialized at $\boldsymbol{x}$, and $\boldsymbol{y}_{[-i]}$ denotes a subset of $\boldsymbol{y}$ with y_i being deleted.

We fitted the very-soft-core model to the Spanish town data shown in Figure 4(b). The data consists of $n = 69$ points in an area of 40×40 square miles. This dataset has been analyzed by Ripley (1977), Ogata and Tanemura (1984) and Gu and Zhu (2001) using the same model. The double MH sampler was applied to this example. In the Metropolis-within-Gibbs cycle, each y_i is updated through 10 consecutive MH steps with a random walk proposal $N(y_i, 5^2 I_2)$. The sampler was run 5 times independently. Each run consisted of 20500 iterations, where the first 500 iterations were discarded for the burn-in process, and 4000 samples were collected from the remaining iterations at equally spaced time points. The overall acceptance rate of the double MH moves was 0.27, and the CPU time cost by each run was about 266s. Averaging over the estimated obtained from the five runs, we got an estimate $\widehat{\theta} = 0.176$ with the standard error 0.001. This estimate is consistent with the Monte Carlo MLE 0.167 obtained by Gu and Zhu (2001), but the latter has a large standard error of 0.078.

5 Discussion

We have proposed the double MH sampler for conducting an approximate Bayesian analysis for the models with intractable normalizing constants. The double MH sampler removes the need of exact sampling, the auxiliary variables being generated using MH kernels, and thus can be applied to a wide range of problems for which exact sampling is not available or very expensive. Besides the spatial models studied in the paper, the double MH sampler can be directly applied to many other scientific models or problems, such as image segmentation (see e.g., Hurn *et al.*, 2003), social network modeling (see e.g., Wasserman and Fraust, 1994), and genetic analysis (Francois *et al.*, 2006).

As a practical hint, we would like to point out that the MCMC sampler used for generating auxiliary variables is the key to the efficiency of the double MH sampler. In this paper, we used the Gibbs sampler for the first and second examples, and used the Metropolis-within-Gibbs sampler for the third example. Theoretically, any MCMC samplers can be used here, and some may be even more efficient than the ones we used. For example, the Swendsen-Wang algorithm (Swendsen and Wang, 1987) or the Wolff algorithm (Wolff, 1989) can be used for the autologistic models, and they are expected to be more efficient than the plain Gibbs sampler we used. Similarly, the block Gibbs sampler (Liu *et al.*, 1995) can be used for the spatial point process with the points being grouped appropriately.

When the dimension of the problem is high, the curse of dimensionality may be a serious difficulty for the MH sampler. To maintain a given level of quality of auxiliary variables, its number of iterations need to increase exponentially with dimension. In this case, the sequential parallel tempering algorithm (Liang, 2003) may be used. As demonstrated in Liang (2003), the sequential parallel tempering algorithm can significantly reduce the curse of dimensionality suffered by the MH sampler.

Acknowledgments

The author's research was partially supported by the grant (DMS-0607755) made by the National Science Foundation and the award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST).

References

Andrews, D.F. and Herzberg, A.M. (1985). *Data.* New York: Springer.

Balram, N. and Moura, J.M.F. (1993). *Noncausal Gauss Markov random fields: Parameter structure and estimation. IEEE Trans. Inform. Theory*, **39**, 1333-1355.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. B*, 36, 192-236.

Besag, J.E. and Moran, P.A.P. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, **62**, 555-562.

Childs, A.M., Patterson, R.B. and MacKay, D.J.C. (2001). Exact sampling from nonattractive distributions using summary states. *Phys. Rev. E.* **63**, 036113.

Dryden, I., Ippoliti, L., and Romagnoli, L. (2002). Adjusted maximum likelihood and pseudo-likelihood estimation for noisy Gaussian Markov random fields. *J. Comput. Graph. Statist.*, **11**, 370-388.

Francois, O., Ancelet, S. and Guillot, G. (2006). Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**, 805-816.

Gelman, A., Roberts, R.O. and Gilks, W.R. (1996). Efficient Metropolis Jumping rules. In *bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, new York: Oxford University Press.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457-472.

Geyer, C.J. (1991). Markov chain Monte Carlo maximum likelihood, in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, ed. E.M. Keramigas, Fairfax: Interface Foundation, pp.156-163.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. R. Statist. Soc. B*, 54, 657-699.

Green, P.J. and Richardson, S. (2002). Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.*, 97, 1055-1070.

Gu, M.G. and Zhu, H. (2001). maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *J. R. Statist. Soc. B*, **63**, 339-355.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation* (2nd edition). Prentice Hall International, Inc.

Huang, F. and Ogata, Y. (1999). Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models. *J. Comput. Graph. Statist.*, **8**, 510-530.

Huang, F. and Ogata, Y. (2002). Generalized pseudo-likelihood estimates for Markov random fields on lattice. *Ann. Inst. Statist. Math.*, **54**, 1-18.

Hukushima K. and Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.*, **65**, 1604-1608.

Hurn, M., Husby, O. and Rue, H. (2003). A tutorial on image analysis. *Lecture Notes in Statistics* **173**, 87-141.

Liang, F. (2003). Use of sequential structure in simulation from high dimensional systems. *Physical Review E*, **67**, 56101-56107.

Liang, F. (2007). Continuous Contour Monte Carlo for Marginal Density Estimation with an Application to a Spatial Statistical Model. *J. Comput. Graph. Statist.* **16**, 608-632.

Liang, F., Liu, C. and Carroll, R.J. (2007). Stochastic Approximation in Monte Carlo Computation. *J. Amer. Statist. Assoc.*, **102**, 305-320.

Møller, J., Pettitt, A.N., Reeves, R., and Berthelsen, K.K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalizing constants. *Biometrika*, **93**, 451-458.

Moran, P.A.P. (1973). A Gaussian Markovian process on a square lattice. *J. Appl. Probab.*, **10**, 54-62.

Müller, P. (1993). Alternatives to the Gibbs sampling scheme. Technical report, Institute pf Statistics and Decision Sciences, Duke University.

Murray, I., Ghahramani, Z. and MacKay, D.J.C. (2006). MCMC for doubly-intractable distributions. *Proc. 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI).*

Ogata, Y. and Tanemura, M. (1984). Likelihood analysis of spatial point patterns. *J. R. Statist. Soc. B*, **46**, 496-518.

Preisler, H.K. (1993). Modeling spatial patterns of trees attacked by bark-beetles. *Appl. Statist.*, **42**, 501-514.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). *Numerical Recipes in C* (2nd edition). Cambridge University Press.

Propp, G. and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struc. Algor.* **9**, 223-252.

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, Washington, D.C.: U.S. Government Printing Office.

Ripley, B.D. (1977). Modeling spatial patterns (with discussion). *J. R. Statist. Soc. B*, **39**, 172-212.

Sherman, M., Apanasovich, T.V. and Carroll, R.J. (2006). On estimation in Binary autologistic spatial models. *J. Statist. Comput. Simu.*, **76**, 167-179.

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

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

Wolff, U. (1989). Collective Monte Carlo updating for spin systems. *Physical Review Letters*, **62**, 361-364.

Wu, H. and Huffer, F.W. (1997). Modeling the distribution of plant species using the autologistic regression model. *Ecological Statistics*, **4**, 49-64.