Continuous Contour Monte Carlo for Marginal Density Estimation With an Application to a Spatial Statistical Model

Faming LIANG

The problem of marginal density estimation for a multivariate density function $f(\mathbf{x})$ can be generally stated as a problem of density function estimation for a random vector $\lambda(\mathbf{x})$ of dimension lower than that of $\mathbf{x}$. In this article, we propose a technique, the so-called continuous Contour Monte Carlo (CCMC) algorithm, for solving this problem. CCMC can be viewed as a continuous version of the contour Monte Carlo (CMC) algorithm recently proposed in the literature. CCMC abandons the use of sample space partitioning and incorporates the techniques of kernel density estimation into its simulations. CCMC is more general than other marginal density estimation algorithms. First, it works for any density functions, even for those having a rugged or unbalanced energy landscape. Second, it works for any transformation $\lambda(\mathbf{x})$ regardless of the availability of the analytical form of the inverse transformation. In this article, CCMC is applied to estimate the unknown normalizing constant function for a spatial autologistic model, and the estimate is then used in a Bayesian analysis for the spatial autologistic model in place of the true normalizing constant function. Numerical results on the U.S. cancer mortality data indicate that the Bayesian method can produce much more accurate estimates than the MPLE and MCMLE methods for the parameters of the spatial autologistic model.

Key Words: Autologistic models; Kernel density estimation; Reversible jump Markov chain Monte Carlo; Stochastic approximation; Wang-Landau algorithm.

1. INTRODUCTION

A common computational problem in statistics is the calculation of marginal densities. When the joint density is too complicated to allow for an analytical solution, we must turn to approximations. Such a situation arises naturally in the problems of Bayesian model selection and spatial statistical model estimation. The goal of the former problem is to estimate the Bayes factors for a set of prespecified models; that is, to estimate a marginal mass function defined on a set of points. This problem has been tackled by many authors using a

Faming Liang is Associate Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: *fliang@stat.tamu.edu*).

Journal of Computational and Graphical Statistics, Volume 16, Number 3, Pages 608–632
DOI: 10.1198/106186007X238459

variety of approaches including reversible jump MCMC (Green 1995; Green and Richardson 2002), ratio importance sampling (Chen and Shao 1997a,b), path sampling (Meng and Wong 1996; Gelman and Meng 1998), reverse logistic regression (Geyer 1994), marginal likelihood (Chib 1995; Chib and Jeliazkov 2001; Ishwaran, James, and Sun 2001), among others. The latter problem can be described as follows. Let $f(s|\boldsymbol{\theta}) = \psi(\mathbf{s}, \boldsymbol{\theta})/\varphi(\boldsymbol{\theta})$ denote the probability mass function of a spatial statistical model, where $\mathbf{s}$ denotes a configuration of the model, and and $\boldsymbol{\theta}$ the parameter vector of the model. For some spatial statistical models, for example, Ising, autologistic, autonormal, and very-soft-core models (Ripley 1981), the normalizing constant function $\varphi(\boldsymbol{\theta})$ is not available analytically. Evaluating $\varphi(\boldsymbol{\theta})$ can be treated as a problem of marginal density estimation by viewing $\psi(\mathbf{s}, \boldsymbol{\theta})$ as an unnormalized joint distribution of $\mathbf{s}$ and $\boldsymbol{\theta}$.

This article will focus on the latter problem, which can be stated in a general form as follows. Let $\mathbf{x}$ denote a d-dimensional random vector, and let $f(\mathbf{x})$ denote its density function which is known up to a normalizing constant; that is,

$$f(\mathbf{x}) = \frac{1}{C}\psi(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}, \tag{1.1}$$

where $\mathcal{X}$ is the sample space, C is the unknown normalizing constant, and $\psi(\mathbf{x})$is fully known or at least calculable for any points in $\mathcal{X}$. Let $\mathbf{y} = \lambda(\mathbf{x})$ denote an arbitrary function which maps $\mathcal{X}$ to a lower dimensional space $\mathcal{Y}$ with dimension $d_\lambda < d$. The goal is to estimate the marginal density $\xi(\mathbf{y}) = \int_{\{\mathbf{x}:\mathbf{y}=\lambda(\mathbf{x})\}} f(\mathbf{x})d\mathbf{x}$ for $\mathbf{y} \in \mathcal{Y}$. To this end, approximate samples can often be generated from $f(\mathbf{x})$ via MCMC samplers, and the marginal density can then be estimated using the kernel density estimation method (e.g., Wand and Jones 1995). The kernel density estimation method allows for dependent samples (Hart and Vieu 1990; Yu 1993; Hall, Lahiri, and Truong 1995). When independently and identically distributed samples are available, other nonparametric density estimation methods, such as local likelihood (Loader 1999), smoothing spline (Gu 1993; Gu and Qiu 1993), and logspline (Kooperberg and Stone 1991; Kooperberg 1998), are also applicable to estimate the marginal density.

The problem has also been tackled by some authors from a different angle. For example, Chen (1994) proposed an importance sampling-based parametric method for the case where $\mathbf{y}$ is a subvector of $\mathbf{x}$. Chen's estimator also allows for dependent samples. The major shortcoming with Chen's methods is its strong dependence on the knowledge of the analytical form of the inverse transformation $\mathbf{x} = \lambda^{-1}(\mathbf{y})$. Other parametric methods (Gelfand, Smith, and Lee 1992; Verdinelli and Wasserman 1995) suffer from similar handicaps.

The aforementioned methods have a common feature: they are all sample-based. Hence, they will fail to produce an estimate when identically distributed samples (including MCMC samples) cannot be generated from $f(\mathbf{x})$. The difficulty is two-fold: $f(\mathbf{x})$ has a rugged energy landscape on $\mathcal{X}$ and $f(\mathbf{x})$ has an unbalanced energy landscape on different regions of $\mathcal{X}$. The function, $-\log \psi(\mathbf{x})$, is known in statistical physics as the energy function of $f(\mathbf{x})$. In the first case, the energy landscape of the distribution contains a multitude of local energy minima separated by high energy barriers. In simulation from such a distribution, conventional MCMC samplers, such as the Metropolis-Hastings (MH) algorithm (Metropolis et al. 1953; Hastings 1970) and the Gibbs sampler (Geman and Geman 1984), tend to get

trapped in a local energy minimum indefinitely, rendering the simulation ineffective. This difficulty can be alleviated to some extent by employing an advanced MCMC sampler, such as parallel tempering (Geyer 1991), evolutionary Monte Carlo (Liang and Wong 2001), and the slice sampler (Neal 2003). In the second case, MCMC samples can be drawn only from the low-energy region of $\mathcal{X}$. This difficulty cannot be alleviated by employing advanced MCMC samplers as explained at the end of Section 5.

In this article, we propose a new algorithm for estimating marginal densities. The new algorithm can be viewed as a continuous version of the contour Monte Carlo (CMC) algorithm proposed by Liang (2005, 2006). Henceforth, the new algorithm will be abbreviated as the continuous CMC algorithm or CCMC. CCMC is very general. It works for any transformation $\lambda(\mathbf{x})$ regardless of the availability of the analytical form of the inverse transformation. Like CMC, CCMC has the capability to self-adjust the acceptance probability of local moves. This mechanism enables it to escape from local energy minima to sample relevant parts of the sample space very quickly. As illustrated by our examples studied in Sections 4 and 5, respectively, CCMC works for both types of density functions, those having a rugged energy landscape and those having an unbalanced energy landscape.

The remaining part of this article is organized as follows. In Section 2, we briefly review the CMC algorithm. In Section 3, we describe the CCMC algorithm and prove a theorem concerning its convergence. In Section 4, we apply CCMC to a mixture distribution example, which illustrates the performance of CCMC for a distribution with a rugged energy landscape. In Section 5, we apply CCMC to a spatial autologistic model, which illustrates the performance of CCMC for a distribution with an unbalanced energy landscape. In Section 6, we conclude the article with a brief discussion on the use of CCMC in some statistical problems other than spatial statistical models.

2. A BRIEF REVIEW FOR CONTOUR MONTE CARLO

The Wang-Landau (WL) algorithm (Wang and Landau 2001) is a dynamic Monte Carlo algorithm used to calculate the spectral density for a physical system. A remarkable feature of the WL algorithm is that it is not trapped by local energy minima. This feature has led to many successful applications of the algorithm in statistical physics; however, the algorithm has a shortcoming in convergence, and the estimates produced by it can reach only a limited statistical accuracy. The CMC algorithm proposed by Liang (2006) can be regarded as a stochastic approximation correction of the WL algorithm. The CMC algorithm overcomes the shortcoming of the WL algorithm in convergence; and the estimates produced by it can be improved continuously as the simulation proceeds. The CMC algorithm can be described as follows.

Suppose that we are working with the distribution (1.1) and that the sample space $\mathcal{X}$ has been partitioned into the following m disjoint subregions according to the function $\lambda(\mathbf{x})$: $E_1 = \{\mathbf{x} : \lambda(\mathbf{x}) \le \lambda_1\}$, $E_2 = \{\mathbf{x} : \lambda_1 < \lambda(\mathbf{x}) \le \lambda_2\}$, $\ldots$, $E_{m-1} = \{\mathbf{x} : \lambda_{m-2} < \lambda(\mathbf{x}) \le \lambda_{m-1}\}$, and $E_m = \{\mathbf{x} : \lambda(\mathbf{x}) > \lambda_{m-1}\}$, where $\lambda_1, \ldots, \lambda_{m-1}$ are $m-1$ known real numbers. Note that $\lambda(\mathbf{x})$ can be any functions of interest to us. For example, if $f(\mathbf{x})$ is a likelihood function with $\mathbf{x}$ being the parameter vector, and we are interested in finding the

MLE of the parameters, we can set $\lambda(\mathbf{x}) = -\log \psi(\mathbf{x})$. For simplicity, we here assume that the probability value carried by each subregion is nonzero, that is, $\int_{E_i} f(x)dx > 0$ for all $i = 1, \ldots, m$, but CMC allows for the existence of zero-probability subregions.

To facilitate sampling from different subregions, a different weight is attached to each subregion. Let $g_i = \int_{E_i} f(\mathbf{x})d\mathbf{x}/\pi_i$, $i = 1, \ldots, m$, denote the respective weights attached to the m subregions, where $\pi_1, \ldots, \pi_m$ are m prespecified positive real numbers satisfying the constraint $\sum_{i=1}^m \pi_i = 1$. Define

$$f^*(\mathbf{x}) \propto \sum_{i=1}^m \frac{\psi(\mathbf{x})}{g_i} I(\mathbf{x} \in E_i),$$

where $I(\cdot)$ is the indicator function. It is easy to see that sampling from $f^*(\mathbf{x})$ will lead to a random walk in the space of subregions (by regarding each subregion as a "point") with each subregion being sampled with a frequency proportional to π_i. In this sense, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ is called the desired sampling distribution of the subregions. CMC provides a dynamic procedure to learn the weights $g_1, \ldots, g_m$ simultaneously.

Let $\widehat{g}_i^{(t)}$ denote the working estimate of g_i obtained at iteration t; let

$$\widehat{f}^{(t)}(\mathbf{x}) \propto \sum_{i=1}^m \frac{\psi(\mathbf{x})}{\widehat{g}_i^{(t)}} \pi_i I(\mathbf{x} \in E_i) \tag{2.1}$$

denote the working density at iteration t; let $\mathbf{x}_k^{(t)}$, $k = 1, \ldots, M$, denote the samples drawn from $\widehat{f}^{(t)}(\mathbf{x})$; and let $\boldsymbol{\nu}^{(t)} = (\nu_1^{(t)}, \ldots, \nu_m^{(t)})$ denote the realized sampling frequency of the m subregions with $\nu_i^{(t)} = \frac{1}{M}\sum_{k=1}^M I(\mathbf{x}_k^{(t)} \in E_i)$. One iteration of CMC consists of two steps.

CMC algorithm

(a) **(Sampling)** Draw samples $\mathbf{x}_k^{(t)}$, $k = 1, \ldots, M$, from the working density $\widehat{f}^{(t)}(\mathbf{x})$ via a MCMC sampler, e.g., the MH algorithm or the Gibbs sampler.

(b) **(Weight updating)** Update the working estimate $\widehat{\mathbf{g}}^{(t)} = (\widehat{g}_1^{(t)}, \ldots, \widehat{g}_m^{(t)})$ recursively by setting

$$\log \widehat{g}_i^{(t+1)} = \log \widehat{g}_i^{(t)} + \delta_t \left(\nu_i^{(t)} - \pi_i\right), \quad i = 1, \ldots, m, \tag{2.2}$$

where δ_t is called the gain factor.

Note that in the initial iteration, we have $t = 0$ and $\log \widehat{g}_1^{(0)} = \cdots = \log \widehat{g}_m^{(0)} = 0$. The gain factors are positive and nonincreasing, and satisfy the conditions

$$\sum_{t=0}^{\infty} \delta_t = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \delta_t^2 < \infty, \tag{2.3}$$

where the first condition ensures that the solution can be reached from any initial points when the number of iterations is large enough, and the second condition asymptotically damps the

effect of the random errors introduced by the innovation $\nu_i^{(t)}$. Refer to a classical stochastic approximation book, for example, Nevelson and Hasminskii (1973), for more explanations of the conditions. There are many ways to choose the sequence $\{\delta_t : t = 1, 2, \dots\}$ to satisfy (2.3). For example, Liang (2006) suggested the following,

$$\delta_t = \frac{\rho\kappa}{\max(\kappa, t)}, \quad \kappa > 0,\ t = 0, 1, 2, \dots, \tag{2.4}$$

for a given value of κ. In theory, the choice of κ should not affect the convergence of the algorithm; however, in practice, a good choice of κ may affect the stability of the algorithm. Also, a large value of κ will enable the sampler to reach all subregions very quickly even for a large system.

Based on the standard theory of stochastic approximation (Robbins and Monro 1951; Blum 1954), Liang (2006) proved that,

$$P\left\{\lim_{t\to\infty} \log \widehat{g}_i^{(t)} = c + \log\left(\int_{E_i} \psi(\mathbf{x})d\mathbf{x}\right) - \log(\pi_i)\right\} = 1, \quad i = 1, \dots, m, \tag{2.5}$$

where c is an arbitrary constant which can be determined by imposing an additional constraint on $\widehat{\mathbf{g}}^{(t)}$. Hence, when t becomes large, $\widehat{g}_i^{(t)} \propto \int_{E_i} \psi(\mathbf{x})d\mathbf{x}/\pi_i$ holds approximately, and sampling from $f^{(t)}(\mathbf{x})$ is approximately equivalent to sampling from $\boldsymbol{\pi}$ if each subregion is viewed as a point. The proof is based on the assumption that the samples drawn in Step (a) are exactly from $\widehat{f}^{(t)}(\mathbf{x})$. However, this assumption can be relaxed to that $\nu_i^{(t)}$ satisfies the condition

$$E\left(\nu_i^{(t)}\right) = S_i^{(t)}/S^{(t)}, \quad i = 1, \dots, m, \tag{2.6}$$

where $S_i^{(t)} = \int_{E_i} \psi(\mathbf{x})d\mathbf{x}/\widehat{g}_i^{(t)}$ and $S^{(t)} = \sum_{j=1}^m S_j^{(t)}$.

To ensure the validity of (2.6), M does not need to be a large number. This can be argued as follows. When t becomes large, sampling from $f^{(t)}(\mathbf{x})$ is approximately equivalent to sampling from $\boldsymbol{\pi}$. Since $\boldsymbol{\pi}$ is usually chosen to be a distribution that can be easily mixed by a MCMC sampler, for example, the uniform distribution, M is not required to be large at this stage. When t is small, the validity of (2.6) is not a concern, since the goal of the early stage simulations is just to generate starting values for the latter stage simulations. The self-adjusting mechanism of CMC, which adjusts the sampling probability of E_i (by adjusting the value of $\widehat{g}_i^{(t)}$) in the adverse direction of the realized sampling frequency of E_i, guarantees the success of the "initialization" process.

3. CONTINUOUS CONTOUR MONTE CARLO

Despite its success, CMC does not work well for density estimation problems because it only provides a histogram estimate for the marginal density $\xi(\mathbf{y})$. In this article we provide a continuous version for CMC. CCMC abandons the use of sample space partitioning and incorporates the technique of kernel density estimation into simulations. Consequently, this incorporation improves the convergence of the simulation.

In what follows, CCMC is described for the case where $\mathbf{y}$ is a bivariate vector. Let $\xi(\mathbf{y})$ be evaluated at the grid points of a lattice. Without loss of generality, we assume that the endpoints of the lattice form a rectangle denoted by $\mathcal{Y}$. In practice, the rectangle can be chosen such that its complementary space is of little interest to us. For example, in Bayesian statistics, the parameter space is often unbounded, $\mathcal{Y}$ can then be set to a rectangle which covers the high posterior density region. A rough high posterior density region can usually be identified based on a preliminary analysis of the data. Alternatively, the high posterior density rectangle can be identified by a trial-and-error process, starting with a small rectangle and increasing it gradually until the resulting estimates converges.

Let the grid points be denoted by $\{\mathbf{z}_{ij} : i = 1, \ldots, L_1, j = 1, \ldots, L_2\}$, where $\mathbf{z}_{ij} = (z_i^{(1)}, z_j^{(2)})$. Let $d_1 = z_2^{(1)} - z_1^{(1)}$ and $d_2 = z_2^{(2)} - z_1^{(2)}$ be the horizontal and vertical neighboring distances of the lattice, respectively. Let $g_{ij} = \xi(\mathbf{z}_{ij})$ be the true marginal density value at the point $\mathbf{z}_{ij}$, and let $\widehat{g}_{ij}^{(t)}$ be the working estimate of g_{ij} obtained at iteration t. For any point $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2) \in \mathcal{Y}$, the density value can then be approximated by bilinear interpolation as follows. If $z_i^{(1)} < \tilde{y}_1 < z_{i+1}^{(1)}$ and $z_j^{(2)} < \tilde{y}_2 < z_{j+1}^{(2)}$ defines i and j, then

$$\widehat{\xi}^{(t)}(\tilde{\mathbf{y}}) = (1-u)(1-v)\widehat{g}_{ij}^{(t)} + u(1-v)\widehat{g}_{i+1,j}^{(t)} + (1-u)v\widehat{g}_{i,j+1}^{(t)} + uv\widehat{g}_{i+1,j+1}^{(t)}, \quad (3.1)$$

where $u = (\tilde{y}_1 - z_i^{(1)})/(z_{i+1}^{(1)} - z_i^{(1)})$ and $v = (\tilde{y}_2 - z_j^{(2)})/(z_{j+1}^{(2)} - z_j^{(2)})$.

Similarly to the desired sampling distribution used in CMC, we specify $\pi(\mathbf{y})$ as the desired sampling distribution on the marginal space $\mathcal{Y}$. Theoretically, $\pi(\mathbf{y})$ can be any distribution defined on $\mathcal{Y}$. Let $\{\delta_t : t = 1, 2, \ldots\}$ denote a sequence of gain factors as used in CMC, and let $\mathbf{H}^{(t)} = \mathrm{diag}(h_{t1}^2, h_{t2}^2)$ denote the bandwidth matrix used in the density estimation procedure at iteration t. The sequence $\{h_{ti} : t = 1, 2, \ldots\}$ is positive and nonincreasing, and converges to 0 as $t \to \infty$. In this article, we set

$$h_{ti} = \min\left\{\delta_t^{\gamma}, \frac{\mathrm{range}(\tilde{y}_i)}{2\left(1 + \log_2(M)\right)}\right\}, \quad i = 1, 2, \quad (3.2)$$

where $\gamma \in (\frac{1}{2}, 1]$, and the second term in $\min\{\cdot, \cdot\}$ is the default bandwidth used in conventional density estimation procedures, for example, the procedure *density(·)* in S-Plus 5.0 (Venables and Ripley 1999, pp. 135). For convenience, we link the choices of $\{\delta_t\}$ and $\{h_t \cdot\}$ together in this article, although this is not necessary in theory. With the notations defined above, one iteration of CCMC can be described as follows.

CCMC algorithm

(a) **(Sampling)** Draw samples $\mathbf{x}_k^{(t)}$, $k = 1, \ldots, M$, from the working density

$$\widehat{f}^{(t)}(\mathbf{x}) \propto \frac{\psi(\mathbf{x})}{\widehat{\xi}^{(t)}(\lambda(\mathbf{x}))}, \quad (3.3)$$

via a MCMC sampler, for example, the MH algorithm or the Gibbs sampler, where $\psi(\mathbf{x})$ is as defined in (1.1), and $\widehat{\xi}^{(t)}(\lambda(\mathbf{x}))$ is as defined in (3.1).

(b) **(Estimate updating)**

(b.1) Estimate the density of the transformed samples $\mathbf{y}_1^{(t)} = \lambda(\mathbf{x}_1^{(t)}), \ldots, \mathbf{y}_M^{(t)} = \lambda(\mathbf{x}_M^{(t)})$ using the kernel method. Evaluate the density at the grid points,

$$\zeta_u^{(t)}(\mathbf{z}_{ij}) = \frac{1}{M}\sum_{k=1}^{M}\left|\mathbf{H}^{(t)}\right|^{-\frac{1}{2}} K\left((\mathbf{H}^{(t)})^{-\frac{1}{2}}(\mathbf{z}_{ij}-\mathbf{y}_k^{(t)})\right), \quad i=1,\ldots,L_1, \quad j=1,\ldots,L_2, \tag{3.4}$$

where $K(\mathbf{y})$ is a bivariate kernel density function.

(b.2) Normalize $\zeta_u^{(t)}(\mathbf{z}_{ij})$ on the grid points by setting

$$\zeta^{(t)}(\mathbf{z}_{ij}) = \frac{\zeta_u^{(t)}(\mathbf{z}_{ij})}{\sum_{i'=1}^{L_1}\sum_{j'=1}^{L_2}\zeta_u^{(t)}(\mathbf{z}_{i'j'})}, \quad i=1,\ldots,L_1,\ j=1,\ldots,L_2. \tag{3.5}$$

(b.3) Update the working estimate $\widehat{g}_{ij}^{(t)}$ in the following manner,

$$\log\widehat{g}_{ij}^{(t+1)} = \log\widehat{g}_{ij}^{(t)} + \delta_t\left(\zeta^{(t)}(\mathbf{z}_{ij}) - \pi'(\mathbf{z}_{ij})\right), \quad i=1,\ldots,L_1, \quad j=1,\ldots,L_2, \tag{3.6}$$

where $\pi'(\mathbf{z}_{ij}) = \pi(\mathbf{z}_{ij})/[\sum_{i'=1}^{L_1}\sum_{j'=1}^{L_2}\pi(\mathbf{z}_{i'j'})]$.

(c) **(Lattice refinement)** Refine the lattice by increasing the values of L_1, L_2 or both, if $\max\left\{\frac{d_1}{h_{t1}}, \frac{d_2}{h_{t2}}\right\} > \upsilon$, where υ is a threshold value. In this article, we set $\upsilon = 4$ for all examples.

On the convergence of CCMC, we have the following theorem.

Theorem 1. *If $\{\delta_t : t = 1, 2, \ldots\}$ satisfies the condition (2.3) and $\{h_{ti} : t = 1, 2, \ldots\}$ is a sequence as specified in (3.2), then*

$$P\left\{\lim_{t\to\infty}\log\widehat{\xi}^{(t)}(\mathbf{z}_{ij}) = c + \log\xi(\mathbf{z}_{ij}) - \log\pi(\mathbf{z}_{ij})\right\} = 1, \quad i=1,\ldots,L_1, \quad j=1,\ldots,L_2, \tag{3.7}$$

where c is an arbitrary constant which can be determined by imposing an additional constraint on $\widehat{\xi}(\mathbf{z}_{ij})$'s.

Proof: The proof of this theorem mimics a similar proof in the context of CMC. Since $\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_M^{(t)}$ are MCMC samples generated from (3.3), it follows from Wand and Jones (1995, pp. 97) that

$$E\zeta_u^{(t)}(\mathbf{z}_{ij}) = \Psi^{(t)}(\mathbf{z}_{ij}) + \frac{1}{2}\mu_2(K)\mathrm{tr}\left\{\mathbf{H}^{(t)}\mathcal{H}^{(t)}(\mathbf{z}_{ij})\right\} + o\left\{\mathrm{tr}(\mathbf{H}^{(t)})\right\}, \tag{3.8}$$

where $\Psi^{(t)}(\mathbf{z})$ is a density function proportional to $\xi(\mathbf{z})/\widehat{\xi}^{(t)}(\mathbf{z})$, $\mu_2(K) = \int z^2K(z)dz$, and $\mathcal{H}^{(t)}$ is the Hessian matrix of $\Psi^{(t)}(\mathbf{z})$. Thus, $\zeta_u^{(t)}(\mathbf{z})$ forms an (asymptotically) unbiased estimator of $\Psi^{(t)}(\mathbf{z})$ as h_t. goes to 0.

Using Taylor's theorem, we have

$$\begin{aligned} E\zeta^{(t)}(\mathbf{z}_{ij}) &= E\frac{\zeta_u^{(t)}(\mathbf{z}_{ij})}{\sum_{i=1}^{L_1}\sum_{j=1}^{L_2}\zeta_u^{(t)}(\mathbf{z}_{ij})} \approx \frac{\Psi^{(t)}(\mathbf{z}_{ij})}{\sum_{i=1}^{L_1}\sum_{j=1}^{L_2}\Psi^{(t)}(\mathbf{z}_{ij})} \\ &= \frac{\xi^{(t)}(\mathbf{z}_{ij})/\widehat{\xi}^{(t)}(\mathbf{z}_{ij})}{\sum_{i=1}^{L_1}\sum_{j=1}^{L_2}\xi^{(t)}(\mathbf{z}_{ij})/\widehat{\xi}^{(t)}(\mathbf{z}_{ij})}, \end{aligned} \tag{3.9}$$

as $h_{t\cdot}$ goes to 0. The proof can then be completed similarly to Liang (2006) by noting the similarity of (2.6) and (3.9). □

CCMC tends to converge faster than CMC. This is because CCMC employs the techniques of kernel density estimation, and thus the estimates can be updated in large blocks at the early stage of the simulation. In CCMC, the bandwidth plays a similar role as the gain factor. It is meant to enhance the ability of the sampler in sample space exploration. The difference is that the bandwidth controls the moving stepsize of the sampler, while the gain factor controls the moving mobility of the sampler. CCMC is equipped with both the bandwidth and the gain factor.

In the above, CCMC is described in two-dimensional space. Extension to other dimensional spaces, such as one or three-dimensional spaces, is straightforward. For a higher dimensional space, CCMC may be hindered by the curse of dimensionality. First, the multivariate kernel density estimator itself suffers from the curse of dimensionality. High-dimensional space is very different than one-, two-, or three-dimensional space. It is vast, and the points lying in such a space have very few near neighbors. To maintain required estimation accuracy, the sample size should increase exponentially with dimension. Refer to Scott (1992) for more discussions on this issue. Second, CCMC evaluates the kernel density at grid points. As the dimension increases, the number of grid points should increase exponentially, typically, at a rate $O(L_\lambda^d)$, where L denotes the linear size of the lattice.

For an effective implementation of CCMC, several issues need to be addressed.

- *Choice of the desired sampling distribution.* Theorem 1 implies that $\xi(\mathbf{z})/\widehat{\xi}^{(t)}(\mathbf{z})$ converges to $\pi(\mathbf{z})$ when t becomes large. This, together with Equation (3.8), suggests that the uniform distribution is an attractive choice for $\pi(\mathbf{y})$, as it leads to an approximate zero Hessian matrix and thus a small bias of density estimation. In this article, we set $\pi(\mathbf{y})$ to the uniform distribution for all examples. Equation (3.6) can then be simplified to

$$\log \widehat{g}_{ij}^{(t+1)} = \log \widehat{g}_{ij}^{(t)} + \delta_t \zeta^{(t)}(\mathbf{z}_{ij}), \quad i = 1, \ldots, L_1, \quad j = 1, \ldots, L_2, \tag{3.10}$$

  because adding to or subtracting from $\{\log \widehat{g}_{ij}^{(t)}\}$ a constant will not change the dynamics of CCMC.

- *Choice of the MCMC sampler.* As in CMC, the self-adjusting mechanism of CCMC can provide substantial help for the system to escape from local traps. Hence, the choice of the MCMC sampler is no longer crucial to the mixing of CCMC. Our numerical results indicate that the MH algorithm can work well for many problems, even for those having a rugged or severely unbalanced energy landscape. When

the problems are more complicated, an advanced MCMC sampler, such as parallel tempering (Geyer 1991), evolutionary Monte Carlo (Liang and Wong 2001), or the slice sampler (Neal 2003), can be used to avoid possible local-trap problems.

- *Choice of the kernel and bandwidth matrix.* The kernel method is used in CCMC based on two considerations. First, the kernel method allows for the dependence of the MCMC samples (Yu 1993). Second, the asymptotics of (3.8) depend on only the bandwidth matrix, and thus a small value of M is allowed for in (3.8). These two considerations rule out other density estimation techniques, such as local likelihood, logspline, and smoothing spline, although they may have better performance on the boundary regions.

  Before discussing the choices of the kernel and bandwidth matrix, we note that CCMC is different from other density estimators in its dynamic nature. In CCMC, the estimates are updated iteratively, each updating is based on a small number of MCMC samples simulated from the working density, and the updating manner turns gradually from global to local due to the gradual decrease of the kernel bandwidth as iterations progress. Hence, the rules developed for the choices of kernel and bandwidth matrix for conventional kernel density estimators may not work for CCMC.

  In this article, we associate the choice of bandwidth matrix with the choice of $\boldsymbol{\pi}$. Since $\boldsymbol{\pi}$ is set to the uniform distribution, the samples $\mathbf{y}_1^{(t)}, \ldots, \mathbf{y}_M^{(t)}$ tend to be uniformly distributed on $\mathcal{Y}$ (independent of $f(\mathbf{x})$) when t becomes large. This suggests the use of a diagonal bandwidth matrix with the diagonal elements accounting for the variability (range) of the samples in the coordinate directions. Other bandwidth matrices, such as the full bandwidth matrix, are not recommended here due to the uniformity of the samples. When nonequally spaced grid points are used in evaluation of the density, the bandwidth h_{ti} may be allowed to vary with the position of the grid point. However, the data-driven approaches—for example, the nearest neighbor and balloon approaches (Loftsgaarden and Quesenberry 1965; Terrell and Scott 1992)—may not be appropriate to CCMC, because the convergence of bandwidth is out of our control in these approaches.

  As in conventional kernel density estimation problems, the shape of the kernel has little influence on the results. In this article, $K(\cdot)$ is set to the standard d_λ-variate normal density, that is,

$$K(\mathbf{y}) = \frac{1}{(2\pi)^{d_\lambda/2}} \exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{y}\right).$$

  Since the bandwidth tends to zero as t becomes large, this choice also works for the case where the marginal density has some discontinuous points. Refer to Ghosh and Huang (1992) and references therein for methods of estimation of discontinuous densities. Other kernels, for example, the Epanechnikov kernel, should work equally well. It is worth noting that a boundary kernel should work better for CCMC due to the elimination of boundary bias. Refer to Müller and Stadtmüller (1999), Jones (1993), and references therein for development of boundary kernels.

- *Fast computation of kernel estimates*. The kernel method requires us to evaluate the kernel density at all grid points for any given sample $\mathbf{y}_k^{(t)}$. For fast computation, we only need to evaluate the kernel density at the grid points lying in a neighborhood of $\mathbf{y}_k^{(t)}$, for example, the grid points within a cycle of radius $\max\{4h_{t1}, 4h_{t2}\}$ and centered at $\mathbf{y}_k^{(t)}$. This will save a lot of CPU time when the lattice size is large, while keeping the estimation being less affected by the kernel density truncation. Alternatively, the binning techniques developed by Fan and Marron (1994) and Wand (1994) can be used to accelerate the computation of the kernel estimates.

- *Lattice refinement*. Since the normal kernel is used and $1 - \Phi(2)$ is a small number, where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution, a sample lying in the middle of two neighboring grid points will have only very minor contributions to the marginal density estimate when $\min\{d_1/h_{t1}, d_2/h_{t2}\} > 4$. Based on this consideration, we set $\upsilon = 4$ in the step of lattice refinement. In practice, d_1, d_2 and the sequence $\{h_{t\cdot}; t = 1, 2, \dots\}$ are often set a priori such that the lattice does not need to be refined during simulations. Of course, this is done only for convenience. For efficiency of the algorithm, the lattice should be refined sequentially, as this avoids updating $\{\widehat{g}_{ij}\}$ on a large lattice at the early stage of the simulation.

- *Choice of other parameters*. Other parameters of CCMC include the sample size M, the parameters ρ and κ used in specifying $\{\delta_t : t = 1, 2, \dots\}$, the parameter γ used in specifying $\{h_{t\cdot} : t = 1, 2 \dots\}$, and the number of iterations. As argued before, M is not required to be very large for the convergence of CCMC. Based on empirical evidence, Liang (2006) suggested that a value of M between 10 and 100 is appropriate for most problems. In this article, we fixed $M = 10$ in all computations. Our numerical results indicate that this choice is appropriate. The parameters ρ, κ and the number of iterations are related. They can be set according to the following criterion: The more complex the problem is, the larger values they should have. The appropriateness of the settings can be diagnosed based on the deviation of the realized sampling distribution from the desired one. Since we have fixed $\pi(\mathbf{y})$ to the uniform distribution, the transformed samples $\mathbf{y}_1, \mathbf{y}_2, \dots$ should be approximately uniformly distributed on $\mathcal{Y}$ if ρ, κ and the number of iterations are chosen appropriately. If this is not the case, CCMC should be rerun with more iterations, a large value of ρ or a large value of κ.

  The choice of $\{h_{t\cdot} : t = 1, 2, \dots\}$ given in (3.2) implies that the parameter γ controls the smoothness of the resulting marginal density estimate. The smaller γ, the smoother the resulting marginal density estimate will be. However, if γ is too small, the resulting estimate may be over-smoothed and thus biased. Hence, the choice of γ should balance the efficiency of CCMC in sample space exploration and the accuracy of the resulting marginal density estimate. Note that the choice of γ will not affect the convergence of CCMC. For simplicity, we set $\gamma = 1$ in all computations of this article.

4. AN ILLUSTRATIVE EXAMPLE

In this section, we illustrate the use of CCMC as a technique of marginal density estimation through a synthetic example. Our numerical results indicate that CCMC converges much faster than CMC. Consider the following mixture distribution,

$$f(\mathbf{x}) = \frac{1}{3}N_3(-\boldsymbol{\mu}_1, \Sigma_1) + \frac{2}{3}N_3(\boldsymbol{\mu}_2, \Sigma_2), \tag{4.1}$$

where $\mathbf{x} = (x_1, x_2, x_3)$, $\boldsymbol{\mu}_1 = (-5, -5, -5)$, $\boldsymbol{\mu}_2 = (10, 25, 1)$,

$$\Sigma_1 = \begin{pmatrix} 4 & 5 & 0 \\ 5 & 64 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{pmatrix}.$$

Suppose that we want to estimate the marginal density of $\mathbf{y} = (x_1, x_2)$. This example mimics a multimodal posterior distribution in Bayesian inference. The two modes of the distribution are well separated.

CCMC was first applied to this example. We restricted $\mathcal{Y}$ to the square $[-30.5, 30.5] \times [-30.5, 30.5]$, and evaluated the marginal density on a square lattice of size 245×245. We set $M = 10$, $\rho = 20$, $\kappa = 1000$, and the total number of iterations to be 5×10^6. At each iteration, the working density is simulated using the MH algorithm with the random walk proposal $N_3(\mathbf{x}_t, 5^3 I_3)$, where I_3 denotes the 3×3 identity matrix. The overall acceptance rate is about 0.18. Figure 1(a) shows the log-marginal density estimate obtained in one run of CCMC. To show the shape of the first component, the estimate is plotted in the log-scale instead of the true density scale; otherwise, we can only see a flat surface for the first component. Note that for this example, the region $\{\mathbf{x} : \log f(\mathbf{x}) > -16.5\}$ contains more than 99.9% percent of the mass of $f(\mathbf{x})$. Figures 1(c) and (d) compare the contour plots of the estimated and true log-marginal densities. The comparison shows that the estimate is accurate and potentially it can be used in inference for the target distribution in place of the true density. CCMC was then run four more times independently. Table 1 reports the resulting estimates of the probabilities $p_1 = P(x_1 < 0)$ and $p_2 = P(x_2 < 0)$. These estimates are calculated based on the marginal density estimates produced in the five runs.

For comparison, CMC was also applied to this example to estimate the probabilities p_1 and p_2. The sample space was partitioned into 244×244 subregions according to the lattice used by CCMC, the desired sampling distribution was set to the uniform distribution, and the total number of iterations was set to 10^7. Other parameters were set as in CCMC, that is, $M = 10$, $\rho = 20$, and $\kappa = 1{,}000$. The algorithm was run five times independently. The results shown in Table 1 indicate that CMC fails for this example even with twice the number of iterations that CCMC used. Here we use the number of energy evaluations instead of the CPU time to measure the computational cost of a run. Actually, this is a much better measure than the CPU time for comparing efficiency of two algorithms, because the CPU time is usually dominated by the part used for energy evaluations when a complex system is simulated. This measure has long been used in statistical physics (Hesselbo and Stinchcombe 1995).

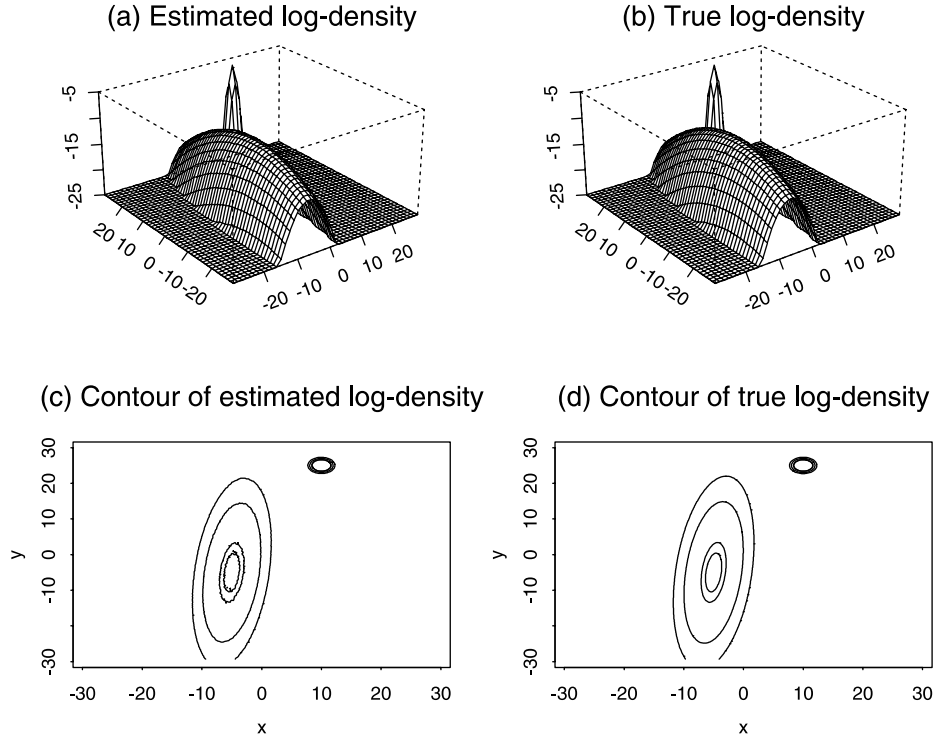

Figure 1. Computational results of CCMC for the mixture Gaussian example. (a) The log-marginal density estimate; (b) the true log-marginal density; (c) the contour plot of the log-marginal density estimate; (d) the contour plot of the true log-marginal density. The circles in the two contour plots corresponds to the 95%, 90%, 50%, and 10% percentiles of the log-density, respectively.

Table 1. Comparison of CMC and CCMC for estimating the probabilities p_1 and p_2. The numbers in the parentheses are the standard deviations of the estimates. The true values of p_1 and p_2 are 0.331 and 0.245, respectively.

Algorithm	p_1	p_2	Number of energy evaluations
CCMC	0.330 (0.003)	0.245 (0.002)	5×10^6
CMC	0.267 (0.137)	0.211 (0.110)	10^7

This example shows that CCMC outperforms CMC in marginal density estimation due to the use of the kernel smoothing techniques. As mentioned before, the kernel smoothing step enhances the ability of CCMC in sample space exploration by penalizing the transitions into the neighborhood of historical samples. One possible way to improve efficiency of CMC is to repartition the sample space according to a coarser lattice, say, a lattice of size 50×50. However, a coarser partition often causes the inability of CMC in transitions between different modes of the distribution, and a rougher histogram estimate for the marginal density.

5. BAYESIAN INFERENCE FOR SPATIAL AUTOLOGISTIC MODELS

5.1 INTRODUCTION

The autologistic model (Besag 1974) has been widely used for spatial data analysis, for example, Preisler (1993), Augustin et al. (1996), Wu and Huffer (1997), and Sherman, Apanasovich, and Carroll (2006). Let $\mathbf{s} = \{s_i : i \in D\}$ denote the observed binary data, where s_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 a set of "neighbors" of spin i. The likelihood function of the model is

$$f(\mathbf{s}|\alpha, \beta) = \frac{1}{\varphi(\alpha, \beta)} \exp\left\{\alpha \sum_{i \in D} s_i + \frac{\beta}{2} \sum_{i \in D} s_i \left(\sum_{j \in N(i)} s_j\right)\right\}, \quad (\alpha, \beta) \in \Omega, \tag{5.1}$$

where Ω is the parameter space, and $\varphi(\alpha, \beta)$ is the normalizing constant defined by

$$\varphi(\alpha, \beta) = \sum_{\text{for all possible } \mathbf{s}} \exp\left\{\alpha \sum_{j \in D} s_j + \frac{\beta}{2} \sum_{i \in D} s_i \left(\sum_{j \in N(i)} s_j\right)\right\}.$$

The parameter α determines the overall proportion of s_i with a value of $+1$, and the parameter β determines the intensity of the interaction between s_i and its neighbors. When β is large, say, near or greater than 0.44 (the critical value of the Ising model), the configuration $\mathbf{s}$ tends to have large clusters of the same orientation, which fluctuate very slowly. The orientation of the cluster is largely determined by the sign of α, and the size of the cluster is determined by the magnitudes of both α and β. When β is large, the Gibbs sampler (Geman and Geman 1984) suffers from an extremely long autocorrelation time in simulating configurations of the model.

A major difficulty with this model is that the normalizing constant is generally unknown analytically and thus the parameters are difficult to estimate. Evaluating $\varphi(\alpha, \beta)$ requires to sum over all $2^{|D|}$ possible realizations of $\mathbf{s}$. An exact evaluation of $\varphi(\alpha, \beta)$ is impossible even for a moderate system. To circumvent this difficulty, many authors approximate the likelihood function by a tractable pseudo-likelihood function, for example, Besag (1974), Besag et al. (1991), Heikkinen and Högmander (1994), and Rydén and Titterington (1998). For

example, Besag (1974) proposed to approximate (5.1) by the following pseudo-likelihood function,

$$PL(\alpha, \beta|\mathbf{s}) = \prod_{i \in D} \frac{\exp\left\{s_i \left(\alpha + \beta \sum_{j \in N(i)} s_j\right)\right\}}{\exp\left\{\alpha + \beta \sum_{j \in N(i)} s_j\right\} + \exp\left\{-\alpha - \beta \sum_{j \in N(i)} s_j\right\}}, \tag{5.2}$$

which avoids the problem of unknown normalizing constant. The estimator maximizing (5.2) is called the maximum pseudo-likelihood estimator (MPLE). When the neighboring dependence is strong, MPLE tends to have a large deviation from the true value. This has been noted by several authors, for example, Huang and Ogata (1997) and Geyer and Thompson (1992).

On the other hand, many authors attempt to find the MLE of the parameters based on Monte Carlo approximations. For example, Geyer and Thompson (1992) proposed the following method. Let (α^*, β^*) be any point in the space Ω, and let $\psi(\alpha, \beta, \mathbf{s}) = \varphi(\alpha, \beta) f(\mathbf{s}|\alpha, \beta)$. Based on the identity

$$\frac{\varphi(\alpha, \beta)}{\varphi(\alpha^*, \beta^*)} = \sum_{\mathbf{s}} \frac{\psi(\alpha, \beta, \mathbf{s})}{\psi(\alpha^*, \beta^*, \mathbf{s})} f(\alpha^*, \beta^*|\mathbf{s}) = E_f \frac{\psi(\alpha, \beta, \mathbf{s})}{\psi(\alpha^*, \beta^*, \mathbf{s})}, \tag{5.3}$$

Geyer and Thompson (1992) proposed to approximate (5.1) by

$$\begin{aligned} L_n(\alpha, \beta|\mathbf{s}) &= \alpha \sum_{i \in D} s_i + \frac{\beta}{2} \sum_{i \in D} s_i \left(\sum_{j \in N(i)} s_j\right) \\ &\quad - \log \varphi(\alpha^*, \beta^*) - \log\left[\frac{1}{n} \sum_{k=1}^{n} \frac{\psi(\alpha, \beta, \mathbf{s}^{(k)})}{\psi(\alpha^*, \beta^*, \mathbf{s}^{(k)})}\right], \end{aligned} \tag{5.4}$$

where $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(n)}$ denote the samples drawn from $f(\mathbf{s}|\alpha^*, \beta^*)$. As $n \to \infty$, $L_n(\alpha, \beta|\mathbf{s})$ approaches $\log f(\mathbf{s}|\alpha, \beta)$. The estimator which maximizes $L_n(\alpha, \beta|\mathbf{s})$ is called Monte Carlo MLE or MCMLE. The same technique has been used by other Monte Carlo based methods. For example, Huang and Ogata (1999) and Gu and Zhu (2001) approximated the first and second derivative of $\log \varphi(\alpha, \beta)$ using averages of Monte Carlo samples. As indicated by the numerical results reported by Gu and Zhu (2001), these methods have similar performances in terms of bias, standard deviation and mean squared errors of the resulting estimates. It is worth noting that the method proposed by Møller et al. (2006) avoids the approximation to the normalizing constant function $\varphi(\alpha, \beta)$ by introducing an auxiliary variable into the MH algorithm for the posterior of the parameters (α, β), and thus is suitable for a Bayesian analysis for the model. However, due to the difficulty in choosing the auxiliary variable distribution, their method only works for the case where both the lattice size and the association parameter β are small.

Finally, we note that some nonsimulation-based methods have been recently proposed for estimating the function $\varphi(\alpha, \beta)$. However, these methods are usually quite restrictive and not applicable to general autologistic models. For example, the method proposed by Pettitt, Friel, and Reeves (2003) only works for a lattice for which each column has two nearest column neighbors and the smallest number of rows and columns is not greater

than 10. The method proposed by Reeves and Pettitt (2004) is exact, but works only for a small rectangular lattice. These two methods will not work for the example studied in the following. Indeed the U.S. cancer mortality data is analyzed using the autologistic model for which the lattice is large ($|D| = 2293$) and has an irregular shape.

5.2 Bayesian Analysis

Before going to the detailed data analysis, we first consider a general Bayesian framework for the analysis of autologistic models. Suppose that (α, β) take values in a compact set $\Omega = [-0.9, 0.9] \times [0, 0.6]$, which covers the strong neighboring dependence region (β can be greater than 0.44) of the model. To conduct the Bayesian analysis, we assume the following prior distribution for (α, β),

$$P(\alpha, \beta) \propto \frac{1}{\varphi^{\tau}(\alpha, \beta)}, \quad (\alpha, \beta) \in \Omega, \tag{5.5}$$

where τ is a hyperparameter. This prior is equivalent to the likelihood of τ independently and identically distributed observations with $\sum_{i \in D} s_i = 0$ and $\sum_{i \in D} s_i (\sum_{j \in N(i)} s_j) = 0$. Hence, it will drive the estimates of α and β to the point (0,0), and thus will be helpful to reduce the correlation of the estimates of α and β. As known by many researchers, when α and β are large, the correlation of the estimates tends to reduce the accuracy of the estimates. This phenomenon can also be observed in Table 3 (p. 626), where the accuracy of MPLE and MCMLE deteriorate as α and β increase. Since, in most practical examples we have only one observation available, τ should be set to a small value. In our example, we set $\tau = 0.0025$ in all computations. A sensitivity analysis has been done below. It shows that the Bayesian estimates are not very sensitive to the choice of τ.

Given an observation $\mathbf{s}$, the posterior distribution of the model can be written as

$$P(\alpha, \beta | \mathbf{s}) \propto \frac{1}{\varphi^{1+\tau}(\alpha, \beta)} \exp\left\{ \alpha \sum_{i \in D} s_i + \frac{\beta}{2} \sum_{i \in D} s_i \left(\sum_{j \in N(i)} s_j \right) \right\}. \tag{5.6}$$

If MCMC samples $(\alpha_1, \beta_1), \ldots, (\alpha_N, \beta_N)$ can be simulated from $P(\alpha, \beta | \mathbf{s})$, then Bayesian inference can be made for the model. However, $\varphi(\alpha, \beta)$ is unknown. The posterior (5.6) can be approximated by

$$\widehat{P}(\alpha, \beta | \mathbf{s}) \propto \frac{1}{R} \sum_{r=1}^{R} \frac{1}{\widehat{\varphi}_r^{1+\tau}(\alpha, \beta)} \exp\left\{ \alpha \sum_{i \in D} s_i + \frac{\beta}{2} \sum_{i \in D} s_i \left(\sum_{j \in N(i)} s_j \right) \right\}, \tag{5.7}$$

where $\widehat{\varphi}_1(\alpha, \beta), \ldots, \widehat{\varphi}_R(\alpha, \beta)$ denote the estimates of $\varphi(\alpha, \beta)$ produced by CCMC in R independent runs. Note that $\varphi(\alpha, \beta)$ can be viewed as a marginal distribution of the unnormalized joint distribution

$$f(\alpha, \beta, \mathbf{s}) \propto \exp\left\{ \alpha \sum_{i \in D} s_i + \frac{\beta}{2} \sum_{i \in D} s_i \left(\sum_{j \in N(i)} s_j \right) \right\}, \quad (\alpha, \beta) \in \Omega, \tag{5.8}$$

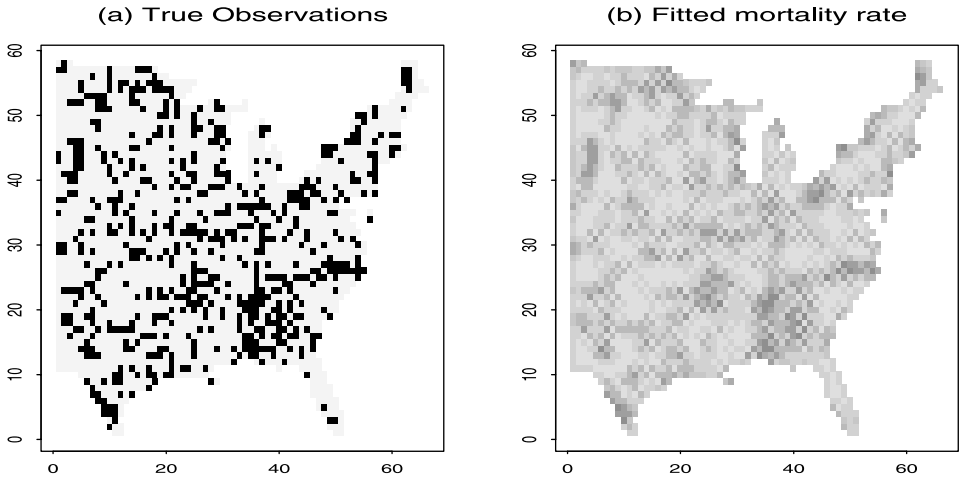

Figure 2. U.S. cancer mortality data. (a) The mortality map of liver and gallbladder cancers (including bile ducts) for white males during the decade 1950–1959. Black squares denote counties of high cancer mortality rate, and white squares denote counties of low cancer mortality rate. (b) Fitted cancer mortality rates by the autologistic model with the parameters being replaced by its Bayes estimates. The cancer mortality rate of each county is represented by the gray level of the corresponding square.

and thus can be estimated using CCMC by setting $\lambda(\mathbf{s}, \alpha, \beta) = (\alpha, \beta)$. Theorem 1 implies that $\widehat{P}(\alpha, \beta|\mathbf{s})$ will converge to the true posterior $P(\alpha, \beta|\mathbf{s})$ almost surely for all $(\alpha, \beta) \in \Omega$ when the run is long enough and both L_1 and L_2 tend to infinity. Let $(\alpha_1, \beta_1), \ldots, (\alpha_N, \beta_N)$ denote MCMC samples drawn from $\widehat{P}(\alpha, \beta|\mathbf{s})$. The standard MCMC theory implies that the Bayesian estimates

$$\widehat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i, \quad \widehat{\beta} = \frac{1}{N}\sum_{i=1}^{N}\beta_i, \tag{5.9}$$

will converge to their respective posterior means as N tends to infinity.

5.3 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 2(a) shows the mortality map of liver and gallblader (including bile ducts) cancers for white males during the decade from 1950–1959. Refer to Sherman, Apamasovich, and Carroll (2006) for more descriptions of the data. The plot shows some apparent geographic clustering.

CCMC was applied to estimate the normalizing constant function $\varphi(\alpha, \beta)$ for this example. Due to its symmetry about α, that is, $\varphi(\alpha, \beta) = \varphi(-\alpha, \beta)$, it only needs to be estimated on the region $[0, 0.9] \times [0, 0.6]$. Considering the boundary bias of kernel density estimation, we estimated $\varphi(\alpha, \beta)$ on a larger region $\mathcal{Y} = [-0.1, 1.0] \times [-0.05, 0.65]$. In simulations, we set $M = 10$, $\rho = 50$, $\kappa = 10{,}000$, and the total number of iterations to be 10^8, and evaluated $\varphi(\alpha, \beta)$ on a rectangular lattice of size 56×71 with steps 0.02 and 0.01

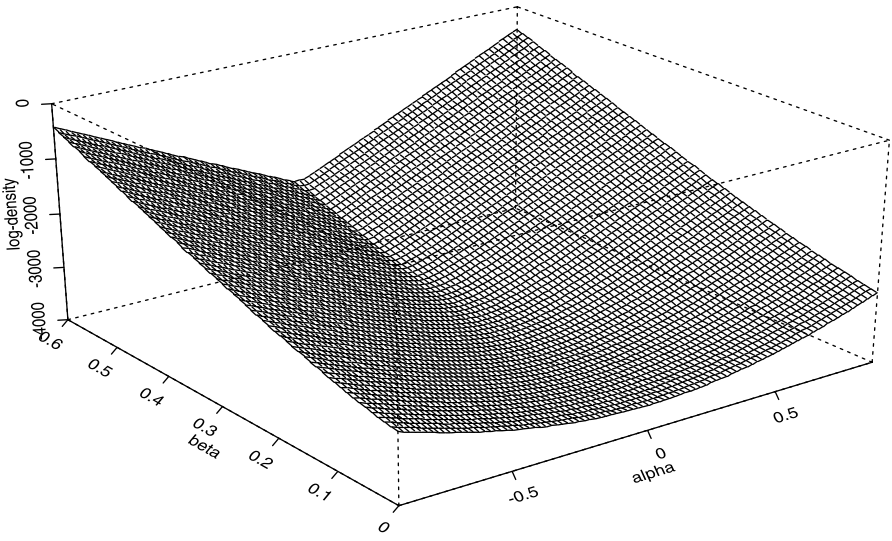

Figure 3. The estimate of $\log \psi(\alpha, \beta)$ produced by CCMC in a single run.

for α and β, respectively. At each iteration, the model configuration $\mathbf{s}$ and the parameters (α, β) are updated alternatively as follows.

(a) (**Sampling**) Repeat (a.1)–(a.3) for $k = 1, \ldots, M$.

(a.1) Draw a random number $u \sim \text{Unif}(0, 1)$.

(a.2) If $u \le \frac{1}{2}$, set $(\alpha_k^{(t)}, \beta_k^{(t)}) = (\alpha_{k-1}^{(t)}, \beta_{k-1}^{(t)})$ and simulate configuration $\mathbf{s}_k^{(t)}$ in a Gibbs iteration cycle (Geman and Geman, 1984) conditional on $(\alpha_k^{(t)}, \beta_k^{(t)})$.

(a.3) If $u > \frac{1}{2}$, set $\mathbf{s}_k^{(t)} = \mathbf{s}_{k-1}^{(t)}$ and simulate $(\alpha_k^{(t)}, \beta_k^{(t)})$ in a single MH move with the invariant distribution proportional to $f(\alpha, \beta, \mathbf{s}_k^{(t)})/\widehat{\xi}^{(t)}(\alpha, \beta)$.

The initial sample is set to the last sample generated at iteration $t-1$; that is, $(\mathbf{s}_0^{(t)}, \alpha_0^{(t)}, \beta_0^{(t)}) = (\mathbf{s}_M^{(t-1)}, \alpha_M^{(t-1)}, \beta_M^{(t-1)})$.

The free boundary condition is assumed for the autologistic model. In Step (a.3), the proposal distribution we adopted for the MH move is a random walk Gaussian proposal $N_2((\alpha_t, \beta_t)', 0.05^2 I_2)$. The overall acceptance rate of the MH moves is about 0.22. CCMC was run five times independently. Figures 4(a) and (b) show the sample paths of α and β obtained in a run. The plots indicate that after the burn-in stage, CCMC can move rather freely in the space $\mathcal{Y}$. Figure 3 shows one estimate of $\varphi(\alpha, \beta)$ produced by CCMC in a run. These estimates were used in all computations of this section.

The MH algorithm was then applied to simulate samples from the approximate posterior (5.7). A total of 5,000 samples were simulated. The resulting estimates of α and β are -0.3008 (4.5e-4) and 0.1231 (2.5e-4), respectively, where the numbers in the parentheses are Monte Carlo errors. These estimates are close to the MPLE estimate ($\widehat{\alpha} = -0.3205$ and $\widehat{\beta} = 0.1115$) and the MCMLE estimate ($\widehat{\alpha} = -0.304$ and $\widehat{\beta} = 0.117$). The MPLE estimate was obtained using the conjugate gradient method, and the MCMLE estimate was obtained

Table 2. Comparison of accuracy of the MPLE, MCMLE, and Bayes estimators for the U.S. cancer mortality example. "SD" stands for the standard deviation of the discrepancy $\mathbf{t}_i^{\text{sim}} - \mathbf{t}_i^{\text{obs}}$, $i = 1, 2$. The values of $\mathbf{t}_i^{\text{sim}}$, $i = 1, 2$, were calculated based on 1,000 independently simulated configurations.

Method	MPLE	MCMLE	Bayes
$(\widehat{\alpha}, \widehat{\beta})$	(−0.3205,0.1115)	(−0.304,0.117)	(−0.3008,0.1231)
$\mathbf{t}_1^{\text{sim}} - \mathbf{t}_1^{\text{obs}}$ $(\times 10^{-2})$	−0.3823	0.6968	−0.0157
SD $(\times 10^{-2})$	0.0708	0.0744	0.0718
$\mathbf{t}_2^{\text{sim}} - \mathbf{t}_2^{\text{obs}}$ $(\times 10^{-2})$	−1.1482	−1.9870	−0.1360
SD $(\times 10^{-2})$	0.1287	0.1291	0.1267

by Sherman, Apanasovich, and Carroll (2006). These estimates indicate an overall majority of low rate counties ($\hat{\alpha} < 0$) and a weak positive correlation between neighboring counties ($\hat{\beta} \approx 0.1$).

To compare the quality of the three estimates, we conduct the following experiment based on the principle of the parametric bootstrap method (Efron and Tibshirani 1993). Let $\mathbf{T}_1 = \sum_{i \in D} s_i / |D|$ and $\mathbf{T}_2 = \frac{1}{2} \sum_{i \in D} s_i \left(\sum_{j \in N(i)} s_j \right) / |D|$. Thus, $\mathbf{T} = (\mathbf{T}_1, \mathbf{T}_2)$ forms a sufficient statistic for the parameters (α, β). Given an estimate $(\widehat{\alpha}, \widehat{\beta})$, we can reversely estimate the quantities $\mathbf{T}_1$ and $\mathbf{T}_2$ by drawing samples from the distribution $f(\mathbf{s}|\widehat{\alpha}, \widehat{\beta})$. If the estimate $(\widehat{\alpha}, \widehat{\beta})$ is accurate, we should have $\mathbf{t}^{\text{obs}} \approx \mathbf{t}^{\text{sim}}$, where $\mathbf{t}^{\text{obs}}$ and $\mathbf{t}^{\text{sim}}$ denote the values of $\mathbf{T}$ calculated from the true and simulated samples, respectively. Table 2 compares the values of $\mathbf{t}^{\text{obs}}$ and $\mathbf{t}^{\text{sim}}$ for the three estimators. To calculate $\mathbf{t}^{\text{sim}}$, 1,000 independent configurations were simulated conditional on each estimate, where each configuration was generated by a run of the Gibbs sampler with a random initial configuration and 1,000 iteration cycles. A convergence diagnostic indicates that 1,000 iterations are long enough for the Gibbs sampler to generate a sample from $f(\mathbf{s}|\widehat{\alpha}, \widehat{\beta})$. Table 2 shows that the values of $\mathbf{t}^{\text{sim}}$ corresponding to the Bayes method are not significant differently from the true values, while this is not the case for MPLE and MCMLE. This experiment indicates that the Bayes method does a better job than MPLE and MCMLE for this example.

To assess the general accuracy of the Bayesian estimator, we simulated 100 independent configurations for the autologistic model under each setting of the parameters (α, β) given in Table 3, and re-estimated the parameters using the Bayes method. The computational results are summarized in Table 3. For comparison, Table 3 also gives the corresponding MPLEs and MCMLEs. The MPLEs and MCMLEs are both calculated using the conjugate gradient method. In computing MCMLEs, (α^*, β^*) is set to the MPLE estimate as in Sherman, Apanasovich, and Carroll (2006), and the sample size n (in Equation (5.4)) is set to 200 with each configuration being generated by a run of the Gibbs sampler with 1,000 iteration cycles. Table 3 shows that the Bayesian method outperforms MPLE and MCMLE in terms of mean squared errors of the resulting estimates. The improvement is especially significant for the strong neighboring dependence region, for example, the points (0,0.4), (0,0.5), (0.4,0.4),

Table 3. Computational results for the simulated data. ave: average of the 100 estimates. sd: standard deviation of the average of the 100 estimates. srmse: square root of mean squared error of the 100 estimates.

		MPLE		MCMLE		Bayes	
(α, β)		$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$
	ave	−0.0053	0.0980	−0.0054	0.0982	−0.0052	0.0972
(0,0.1)	sd	0.0016	0.0015	0.0016	0.0015	0.0016	0.0015
	srmse	0.0171	0.0151	0.0171	0.0150	0.0169	0.0149
	ave	0.0015	0.1987	0.0015	0.1983	0.0016	0.1972
(0,0.2)	sd	0.0013	0.0014	0.0013	0.0012	0.0013	0.0012
	srmse	0.0129	0.0140	0.0129	0.0124	0.0128	0.0125
	ave	0.0003	0.2995	0.0003	0.2973	0.0001	0.2965
(0,0.3)	sd	0.0008	0.0016	0.0008	0.0013	0.0008	0.0013
	srmse	0.0079	0.0159	0.0081	0.0136	0.0079	0.0133
	ave	−0.0011	0.3999	−0.0013	0.3921	−0.0004	0.3966
(0,0.4)	sd	0.0010	0.0016	0.0010	0.0021	0.0004	0.0010
	srmse	0.0097	0.0163	0.0089	0.0195	0.0039	0.0102
	ave	0.0000	0.5023	0.0004	0.4709	0.0002	0.4882
(0,0.5)	sd	0.0049	0.0027	0.0069	0.0032	0.0021	0.0014
	srmse	0.0489	0.0269	0.0438	0.0353	0.0205	0.0182
	ave	0.1007	0.1020	0.1003	0.1015	0.1013	0.1002
(0.1,0.1)	sd	0.0020	0.0016	0.0020	0.0016	0.0020	0.0016
	srmse	0.0199	0.0161	0.0201	0.0156	0.0202	0.0157
	ave	0.1978	0.2018	0.1978	0.2018	0.2007	0.1998
(0.2,0.2)	sd	0.0028	0.0018	0.0029	0.0018	0.0026	0.0017
	srmse	0.0283	0.0183	0.0285	0.0180	0.0255	0.0165
	ave	0.2979	0.3003	0.2981	0.3006	0.3050	0.2972
(0.3,0.3)	sd	0.0063	0.0028	0.0063	0.0028	0.0045	0.0021
	srmse	0.0628	0.0274	0.0627	0.0277	0.0446	0.0215
	ave	0.4188	0.3987	0.4192	0.3994	0.4059	0.4018
(0.4,0.4)	sd	0.0177	0.0058	0.0177	0.0059	0.0114	0.0040
	srmse	0.1773	0.0580	0.1770	0.0588	0.1135	0.0394
	ave	0.5660	0.4860	0.5669	0.4876	0.4923	0.4985
(0.5,0.5)	sd	0.0342	0.0108	0.0342	0.0108	0.0107	0.0038
	srmse	0.3465	0.1079	0.3464	0.1084	0.1069	0.0379

Table 4. Sensitivity analysis for the hyperparameter τ. ave: average of the 100 estimates. sd: standard deviation of the average of the 100 estimates. srmse: square root of mean squared error of the 100 estimates.

		(0,0.2)		(0,0.5)		(0.3,0.3)	
τ		$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\alpha}$	$\widehat{\beta}$
	ave	0.0016	0.1974	0.0002	0.4889	0.3082	0.2964
0.001	sd	0.0013	0.0012	0.0021	0.0014	0.0045	0.0022
	srmse	0.0128	0.0125	0.0205	0.0178	0.0458	0.0217
	ave	0.0016	0.1972	0.0002	0.4882	0.3050	0.2972
0.0025	sd	0.0013	0.0012	0.0021	0.0014	0.0045	0.0021
	srmse	0.0128	0.0125	0.0205	0.0182	0.0446	0.0215
	ave	0.0016	0.1967	0.0002	0.4870	0.2996	0.2984
0.005	sd	0.0013	0.0012	0.0021	0.0014	0.0043	0.0021
	srmse	0.0128	0.0126	0.0205	0.0190	0.0429	0.0211

and (0.5,0.5). The long CCMC run is rewarded with highly accurate Bayesian estimates.

To assess the sensitivity of the Bayesian estimator to the setting of the hyperparameter τ, we conducted a sensitivity analysis for three (α, β) pairs: (0,0.2), (0,0.5), and (0.3,0.3). The results are summarized in Table 4. The estimates only differ at the third digit after the decimal point when τ varies from 0.001 to 0.005 (a five-fold change). This is acceptable to us.

In this example, CCMC is used to estimate the normalizing constant function $\varphi(\alpha, \beta)$ at a set of prespecified parameter points. It seems that the job can also be done by a hybrid MCMC sampler (Müller 1991; Robert and Casella 2004, p. 393) through simulation of the joint distribution

$$g(\alpha', \beta', \mathbf{s}) = \frac{1}{\sum_{(\alpha',\beta')\in\Omega'} \varphi(\alpha', \beta')} \exp\left\{\alpha' \sum_{i\in D} s_i + \frac{\beta'}{2} \sum_{i\in D} s_i \left(\sum_{j\in N(i)} s_j\right)\right\}, \quad (\alpha', \beta') \in \Omega', \tag{5.10}$$

where Ω' denotes the set of prespecified parameter points. For comparison, the hybrid MCMC sampler was applied to estimate $\varphi(\alpha, \beta)$ for this example at the same grid points used by CCMC. Let $(\alpha_t, \beta_t, \mathbf{s}_t)$ denote the state of the Markov chain at iteration t. One iteration of the hybrid sampler consists of the following steps.

(a) Draw a random number $u \sim \text{Unif}(0, 1)$.

(b) If $u \le \frac{1}{2}$, set $(\alpha_t, \beta_t) = (\alpha_{t-1}, \beta_{t-1})$ and simulate $\mathbf{s}_t \sim g(\mathbf{s}|\alpha_t, \beta_t)$ in a Gibbs iteration cycle.

(c) If $u > \frac{1}{2}$, set $\mathbf{s}_t = \mathbf{s}_{t-1}$ and simulate $(\alpha_t, \beta_t) \sim g(\alpha', \beta'|\mathbf{s}_t)$ in a single MH move. The new point is selected at random among the nearest neighboring points of $(\alpha_{t-1}, \beta_{t-1})$.

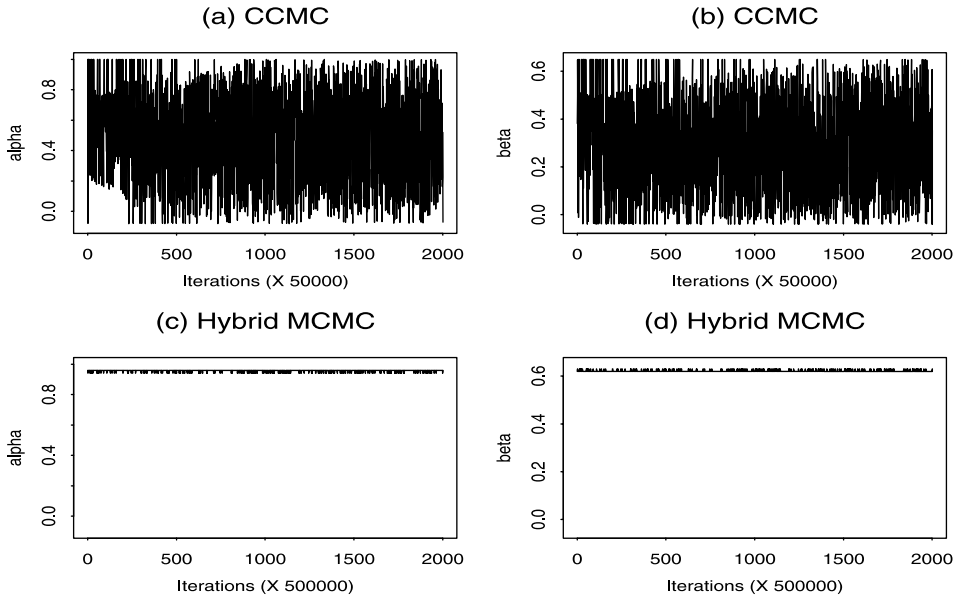

Figure 4. Sample paths of CCMC and the hybrid MCMC sampler for the U.S. cancer mortality example. Plots (a) and (b) show the sample paths of α and β, respectively, obtained in a run of CCMC. Plots (c) and (d) show the sample paths of α and β, respectively, obtained in a run of the hybrid MCMC sampler.

The hybrid sampler started with the point $(-0.1, -0.05)$, moved very quickly to the corner $(1.0, 0.65)$, and then got trapped in the corner never moving out in 10^9 iterations. Note that this run has the same number of energy evaluations as the run of CCMC. The local trap phenomenon can be seen from the sample paths shown in Figures 4(c) and (d). The hybrid sampler fails to sample from relevant parts of the parameter space, and thus fails to produce a good estimate for the function $\varphi(\alpha, \beta)$. Let $\Omega_0 = [0, 1] \times [0, 0.65]$ denote a subset of Ω. In Ω_0, $\varphi(\alpha, \beta)$ has only one mode attained at the corner point (1,0.65). This mode is very high, just like a needle inserted into a flat region. The CCMC estimate tells us that $\max_{(\alpha,\beta)\in\Omega_0} \log \varphi(\alpha, \beta) - \min_{(\alpha,\beta)\in\Omega_0} \log \varphi(\alpha, \beta)$ is approximately 3,109.4. This suggests that producing an accurate estimate for $\varphi(\alpha, \beta)$ has been far beyond the ability of conventional MCMC samplers. Intuitively, to achieve an accurate estimate for $\psi(\alpha, \beta)$ using Monte Carlo samples, the ratio of the number of the samples accumulated from the region around the point (1,0.65) and the number of samples from the region around the point (0,0), which corresponds to the minimum point of $\psi(\alpha, \beta)$, should be approximately e^{3109}. No one can afford such a long run.

The advanced MCMC samplers, such as parallel tempering, evolutionary Monte Carlo or the slice sampler, will also fail for this example due to the limited accuracy of sampling approximation. The strength of the advanced samplers is at transitions between different modes of the distribution instead of sampling from low-density regions. Due to the lack of samples from the low-density region, the nonparametric density estimation methods also fail for this example. Note that the nonparametric density estimation methods belong to the category of sampling frequency approximation based methods. To the best of our

knowledge, CCMC is the first method which can handle such a difficult situation. CCMC approximates the log-marginal density in a dynamic manner.

6. DISCUSSION

In this article, CCMC has been proposed as a general algorithm for marginal density estimation. The generality is reflected in two aspects. First, it works for any transformation $\lambda(\mathbf{x})$, even for one for which the inverse transformation is not available analytically. Second, it works for any distribution, $f(\mathbf{x})$, even for one for which the energy landscape is rugged or unbalanced. When the samples from $f(\mathbf{x})$ are easily available, other methods as reviewed in the introduction can be used. In this case, CCMC may not be the best choice to the problem, because sampling uniformly in the space $\mathcal{Y}$ may cause an unnecessary waste of CPU time.

Besides spatial statistical models, CCMC can potentially be applied for the solution of many other problems provided that the dimension of $\lambda(\mathbf{x})$ is not high. Here are some examples.

Nuisance parameter elimination. Elimination of nuisance parameters is a central problem in statistical inference. Berger, Liseo, and Wolpert (1999) proposed to solve the problem by integrating the joint posterior with respect to the nuisance parameters and then working with the resulting marginal distribution. The marginal likelihood automatically takes into account the uncertainty of the nuisance parameters. Ignoring the uncertainty of the nuisance parameters, for example, the profile likelihood, can cause serious problems when the dimension of the nuisance parameters is high.

Missing data problems. Let $\mathbf{z}_{\text{obs}}$ and $\mathbf{z}_{\text{mis}}$ denote the observed and missed parts of the data, respectively; and let θ denote the parameter vector of the model. Monte Carlo EM (Wei and Tanner 1990) provides a method for finding the MLE of $\boldsymbol{\theta}$ when the conditional expectation in the "E-step" are not available analytically. Otherwise, the EM algorithm (Dempster et al. 1977) can be used. CCMC can work as an alternative to Monte Carlo EM. Once the marginal $f(\theta|\mathbf{z}_{\text{obs}})$ is available numerically, finding the MLE is trivial. One remarkable advantage of this method is that it avoids the local trap problem suffered by EM and Monte Carlo EM.

Bayesian model selection. Although CCMC is developed for continuous random variables, its discrete nature implies that it can be applied to the Bayesian model selection problem as well by setting $\lambda(\mathbf{x})$ to the index of models and representing each model by a different grid point. We note that CMC can also be applied to the Bayesian model selection problem by partitioning the sample space according to the index of models, that is, setting E_i to the parameter space of model M_i. An efficiency comparison of the two methods is of interest. We anticipate that CCMC will outperform CMC because of the use of kernel smoothed estimators in its simulations. As illustrated by the numerical example studied in Section 4, the kernel smoothing step often can improve the convergence of the CMC simulation.

ACKNOWLEDGMENTS

The author thanks Professor L. Tierney, the associate editor, and the referees for their constructive comments which led to significant improvement of this article. The author also thanks Dr. M. Sherman for providing him with the U.S. cancer mortality data, thanks Drs. C. Liu, Y. Lim, and H. Zhu for their early discussions on the project, and thanks Mr. M. Hasson for his help in improving the presentation of this article. The author's research was partially supported by grants from the National Science Foundation (DMS-0405748) and the National Cancer Institute (CA104620).

[Received December 2005. Revised November 2006.]

REFERENCES

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

Berger, J.O., Liseo, B., and Wolpert, R.L. (1999), "Integrated Likelihood Methods for Estimating Nuisance Parameters," *Statistical Science*, 14, 1–28.

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

Besag, J., York, J., and Mollié (1991), "Bayesian Image Restoration with Two Applications in Spatial Statistics," *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Blum, J.R. (1954), "Multidimensional Stochastic Approximation Method," *Annals of Mathematical Statistics*, 25, 737–744.

Chen, M.-H. (1994), "Importance-Weighted Marginal Bayesian Posterior Density Estimation," *Journal of the American Statistical Association*, 89, 818–824.

Chen, M.-H., and Shao, Q.-M. (1997a), "On Monte Carlo Methods for Estimating Ratios of Normalizing Constants," *The Annals of Statistics*, 25, 1563–1594.

——— (1997b), "Estimating Ratios of Normalizing Constants for Densities with Different Dimensions," *Statistica Sinica*, 7, 607–630.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S., and Jeliazkov, I. (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

Dempster, A.P., Laird, N., and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, London: Chapman and Hall.

Fan, J., and Marron, J.S. (1994), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35–56.

Gelfand, A.E., Smith, A.F.M., and Lee, T.M. (1992), "Bayesian Analysis of Constrained Parameter and Truncated Data Problems using Gibbs Sampling," *Journal of the American Statistical Association*, 85, 398–409.

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

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

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, VA: Interface Foundation, pp. 156–163.

——— (1994), "Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo," *Revision of Technical Report No. 568*, School of Statistics, University of Minnesota.

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.

Ghosh, B.K., and Huang, W. (1992), "Optimum Bandwidths and Kernels for Estimating Certain Discontinuous Densities," *Annals of the Institute of Statistical Mathematics*, 44, 563–577.

Green, P.J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Green, P.J., and Richardson, S. (2002), "Hidden Markov Models and Disease Mapping," *Journal of the American Statistical Association*, 97, 1055–1070.

Gu, C. (1993), "Smoothing Spline Density Estimation: A Dimensionless Automatic Algorithm," *Journal of the American Statistical Association*, 88, 495–504.

Gu, C., and Qiu, C. (1993), "Smoothing Spline Density Estimation: Theory," *Annals of Statistics*, 21, 217–234.

Gu, M.G., 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.

Hall, P., Lahiri, S.N., and Truong, Y.K. (1995), "On Bandwidth Choice for Density Estimation with Dependent Data," *Annals of Statistics*, 23, 2241–2263.

Hastings, W.K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and their Applications," *Biometrika*, 57, 97–109.

Hart, J.D., and Vieu, P. (1990), "Data-Driven Bandwidth Choice for Density Estimation Based on Dependent Data," *Annals of Statistics*, 18, 873–890.

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

Hesselbo, B., and Stinchcomble, R.B. (1995), "Monte Carlo Simulation and Global Optimization Without Parameters," *Physical Review Letters*, 74, 2151–2155.

Huang, F., and Ogata, Y. (1997), "Comparison of Two Methods for Calculating the Partition Functions of Various Spatial Statistical Models," *Research Memorandum*, 643, Tokyo: The Institute of Statistical mathematics.

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

Ishwaran, H., James, L.F., and Sun, J. (2001), "Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," *Journal of the American Statistical Association*, 96, 1316–1332.

Jones, M.C. (1993), "Simple Boundary Correction for Kernel Density Estimation," *Statistics and Computing*, 3, 135–146.

Kooperberg, C. (1998), "Bivariate Density Estimation With an Application to Survival Analysis," *Journal of Computational and Graphical Statistics*, 7, 322–341.

Kooperberg, C., and Stone, C.J. (1991), "A Study of Logspline Density Estimation," *Computational Statistics and Data Analysis*, 12, 327–348.

Liang, F. (2005), "Generalized Wang-Landau Algorithm for Monte Carlo Computation," *Journal of the American Statistical Association*, 100, 1311–1327.

——— (2006), "A Theory on Flat Histogram Monte Carlo Algorithms," *Journal of Statistics Physics*, 122, 511–529.

Liang, F., and Wong, W. H. (2001), "Real Parameter Evolutionary Monte Carlo with Applications in Bayesian Mixture Models," *Journal of the American Statistical Association*, 96, 653–666.

Loader, C. (1999), *Local Regression and Likelihood*, New York: Springer.

Loftsgaarden, P.O., and Quesenberry, C.P. (1965), "A Nonparametric Estimate of a Multivariate Probability Density Function," *Annals of Mathematical Statistics*, 28, 1049–1051.

Meng, X.L. and Wong, W.H. (1996) Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, **6**, 831-860.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091.

Møller, J., Pettitt, A.N., Reeves, R., and Berthelsen, K.K. (2006), "An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants," *Biometrika*, 93, 451–458.

Müller, P. (1991), "A Generic Approach to Posterior Integration and Gibbs Sampling," Technical report, Purdue

University, West Lafayette, Indiana.

Müller, H.G., and Stadtmüller, U. (1999), "Multivariate Boundary Kernels and a Continuous Least Squares Principle," *Journal of the Royal Statistical Society*, Series B, 61, 439–458.

Neal, R.M. (2003), "Slice Sampling" (with discussion), *The Annals of Statistics*, 31, 705–767.

Nevelson, M.B., and Hasminskii, R.Z. (1973), *Stochastic Approximation and Recursive Estimation*, American Mathematical Society.

Pettitt, A.N., Friel, N., and Reeves, R. (2003), "Efficient Calculation of the Normalizing Constant of the Autologistic and Related Models on the Cylinder and Lattice," *Journal of the Royal Statistical Society*, Series B, 42, 510–514.

Preisler, H.K. (1993), "Modeling Spatial Patterns of Trees Attacked by Bark-Beetles," *Applied Statistics*, 42, 501–514.

Reeves, R., and Pettitt, A.N. (2004), "Efficient Recursions for General Factorisable Models," *Biometrika*, 91, 751–757.

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. (1981), *Spatial Statistics*, New York: Wiley.

Rydén, T., and Titterington, D.M. (1998), "Computational Bayesian Analysis of Hidden Markov Models," *Journal of Computational and Graphical Statistics*, 7, 194–211.

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, 22, 400–407.

Robert, C.P., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed), New York: Springer.

Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: Wiley.

Sherman, M., Apanasovich, T.V., and Carroll, R.J. (2006), "On Estimation in Binary Autologistic Spatial Models," *Journal of Statistical Computation and Simulation*, 76, 167–179.

Terrell, G.R., and Scott, D.W. (1992), "Variable Kernel Density Estimation," *Annals of Statistics*, 20, 1236–1265.

Venables, W.N., and Ripley, B.D. (1999), *Modern Applied Statistics with S-PLUS* (3rd ed.), New York: Springer.

Verdinelli, I., and Wasserman, L. (1995), "Computing Bayes Factors using a Generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614–618.

Wand, M.P. (1994), "Fast Computation of Multivariate Kernel Estimators," *Journal of Computational and Graphical Statistics*, 3, 433–445.

Wand, M.P., and Jones, M.C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Wang, F., and Landau, D.P. (2001), "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States," *Physical Review Letters*, 86, 2050–2053.

Wei, G.C.G., and Tanner, M.A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85, 699–704.

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

Yu, B. (1993), "Density Estimation in the L^∞ Norm for Dependent Data, with Applications to the Gibbs Sampler," *Annals of Statistics*, 21, 711–735.