

Stochastic Approximation Algorithms for Estimation of Spatial Mixed Models

Hongtu Zhu, Faming Liang, Minggao Gu and Bradley S. Peterson

Abstract

A class of spatial mixed models is introduced first. Spatial mixed models include latent Markov random fields, which make their likelihood functions complex. This complexity in turn makes statistical inferences (e.g., parameter estimates and prediction of latent fields) prohibitively difficult. Therefore, two algorithms are also introduced by integrating recent developments in stochastic approximation algorithms and Monte Carlo methods. The first of these algorithms, a stochastic approximation expectation-maximization (SAEM) algorithm, is developed to estimate the strength of spatial regularization in latent Markov random fields and other parameters. The second algorithm, an annealing stochastic approximation Monte Carlo (ASAMC) algorithm, is proposed to compute optimal estimates of latent fields, which are the global maxima of the likelihood functions of complete data. These algorithms are applied to data sets of the distribution of vegetation species and simulated images to demonstrate their effectiveness.

Keywords: Expectation-maximization; Multicanonical algorithm; Spatial mixed models; Stochastic approximation; Vegetation

1. Introduction

Spatial mixed models (SMM) are natural extensions of generalized linear models and allow for additional components of variability that account for unobservable latent processes. SMMs have wide applications in image analysis, ecology, psychology, physics, and biophysics. For instance, a number of fundamental processes in image analysis, including image restoration, segmentation, and edge-preserving filtering, have been modeled by using SMMs since the seminal work by Besag (1974) and Geman and Geman (1984). See, for example, Zhang (1993), Jalobeanu et al. (2002), Saquib et al. (1998), and Lakshmanan and Derin (1989), among many others. SMMs also include generalized linear mixed models (GLMM) (Breslow and Clayton, 1993; Zhu and Lee, 2002) and latent variable models (LVN) (Bentler and Dudgeon, 1996), both of which can be used to accommodate overdispersion and correlation among outcomes (Zeger et

al., 1988) and to predict and interpolate (or smooth) spatial data (Diggle et al., 1998; Zhang, 2002). Thus, these models have applications in biomedical and educational research, the social sciences, and other fields that investigate complex multivariate, longitudinal, and family data.

However, SMMs are highly complex because of the latent Markov random fields they contain. This complexity makes calculating the maximum likelihood estimates and estimating latent fields prohibitively difficult and therefore poses a major challenge in the applications of SMMs. This challenge can be divided into three distinct issues.

The first issue is that likelihood functions of observed data are often represented by high-dimensional integrals in order to account for latent variables, and these integrals may become intractably complicated. Most of the existing procedures for locating maxima of likelihood functions include the expectation-maximization (EM) algorithm (Dempster et al., 1977), the Monte Carlo EM algorithm (Wei and Tanner, 1990; Booth and Hobert, 1999), Monte Carlo Newton-Raphson algorithm (McCulloch, 1997), and stochastic approximation algorithm (Gu and Kong, 1998; Delyon et al., 1999). However, these optimization algorithms only work for some SMMs (such as GLMMs and LVMs) but not for others, because they strongly depend on a simple likelihood function of complete data (including both latent variables and observed data).

Second, latent Markov random fields (MRF) also add to the complexity of the likelihood functions of SMMs. MRFs have been recently used in a range of fields, such as ecology and image processing, to model spatial and geographical correlation among observations (Winkler, 1995; Li, 2001). For example, ecologists use MRFs to describe the spatially correlated distribution of single or multiple species within a given geographical area (Huffer and Wu, 1998; He et al., 2003). Unknown parameters of MRFs in SMMs usually control the granularity of latent fields, and the corresponding normalizing factor (or partition function) of MRFs, as a function of these unknown control parameters, is not known analytically. In practice, the values of these unknown control parameters are either arbitrarily set or heuristically tuned to particular datasets; maximum likelihood estimates (MLE) of control parameters for MRFs in SMMs are rarely calculated because of the considerable computational burden that is involved. Moreover, most existing optimization algorithms for computing MLEs (Ceyer and Thompson, 1992; Gu and Zhu, 2001) are based on observed MRFs and therefore are inappropriate for latent MRFs in SMMs. Several approximation methods, such as mean-field approximation, have been used to find approximate estimates of control parameters for latent MRFs (Jalobeanu et al., 2002; Qian and Titterton, 1991; Vasconcelos and Lippman, 2001). Recently, a stochastic approximation EM algorithm has been proposed, and its convergence has been established under some conditions (Zhu et al., 2005a, 2005b). However, performance of this SAEM algorithm when applied to many important applications, such as distributions of vegetation data in ecology and imaging analysis, has not yet been investigated.

The third issue in the use of SMMs is how to find optimal estimates of latent fields. This is the central issue of many research questions in ecology, psychology, and image analysis that involve the prediction of latent variables within a data set. For instance, latent field in image segmentation is a set of labels that represents the identities of individual voxels/pixels. In some cases, estimating latent fields is equivalent to mini-

mizing/maximizing a complicated energy function with a large number of variables and is therefore nearly infeasible computationally. Several optimization methods, such as the iterated conditional modes (ICM), can only give a local optimal solution. Stochastic algorithms, such as the simulating annealing and genetic algorithms, have been proposed to search for the globally optimal estimates of latent fields (Kirkpatrick et al., 1983; Holland, 1975); however, these stochastic algorithms converge very slowly and have a high probability of missing the global minimum (Liang, 2005c). Recently, advanced Monte Carlo algorithms, including annealing stochastic approximation and contour Monte Carlo, have been proposed that are efficient for complex simulation and optimization (Liang, 2004, 2005a, 2005b, 2005c). We will apply the annealing stochastic approximation Monte Carlo (ASAMC) algorithm to find optimal latent fields by maximizing the complete-data likelihood functions given MLEs.

In this paper, we formally introduce SMMs and discuss some examples in the fields of ecology. We then propose two advanced stochastic approximation algorithms (SAEM and ASAMC) for calculating MLE and optimal estimates of latent fields in SMMs. Finally, we evaluate the performance of these algorithms using real-world examples, including distributions of vegetation species and image restoration. Throughout the discussion, we will address the three computational issues of SMMs discussed above.

2. Spatial mixed models

We consider stochastic processes $f = \{f(s) : s \in S\}$, $\mathbf{X} = \{X(s) : s \in S\}$, and $\mathbf{Y} = \{Y(s) : s \in S\}$, where $S = \{s : i = 1, \dots, n\}$ is a known discrete index set in R^d . We define SMMs as follows:

- (i) conditional on (f, x) , the components of \mathbf{Y} are mutually independent, and the conditional density of $Y(s)$ given (f, x) is $p^{(Y(s))}(f, x; \omega)$, where ω is an unknown parameter vector;
- (ii) latent field $f = \{f(s) : i = 1, \dots, n\}$ is said to be an MRF with respect to a neighborhood system $\mathcal{N} = \{N_i : i = 1, \dots, n\}$, which is characterized by a Gibbs distribution:

$$p(f|\tau) = \exp\{-U(f, \tau) - \log C(\tau)\}, \quad (1)$$

where $U(f, \tau)$ is a potential (or energy) function, which exhibits the interaction between components of f (Besag, 1974). In addition, the normalizing constant $C(\tau)$ is a partition function having the form

$$C(\tau) = \int_{S_f} \exp\{-U(f, \tau)\} m(df), \quad (2)$$

where S_f is the minimal sample space of f and $m(df)$ is either the Dirac's delta measure or df according to whether f takes discrete or continuous values, respectively.

The above SMMs include many statistical models as special cases. For instance, GLMM is a special class of the SMMs (Breslow and Clayton, 1993) in which f are

random effects. For linear LY models, we have $\mu(s) = \text{E}[y(s)|f, \mathbf{x}] = \mu + \Lambda f(s)$ with f following a multivariate normal distribution (Bentler and Dudgeon, 1996), where Λ is a factor loading matrix. SMMLs also include more general LY models (Lee and Zhu, 2000, 2002).

Let us study two examples from image analysis: image restoration and segmentation.

EXAMPLE 1 (Image restoration). Let s be a pixel-site (or line-site) in a pixelated image, f the true scene and y the observed image, which is a noisy version of f . SMMLs have been used to characterize image construction and restoration. A particular example for image restoration is defined by

$$y = Hf + \varepsilon, \quad (3)$$

where H is the convolution matrix and $\varepsilon \sim N(\mathbf{0}, \phi^{-1}I_n)$, in which I_n is an identity matrix. In this case, we have $\mu(s) = \text{E}[y(s)|f] = H(s)f$. Furthermore, we will assume that the true image f follows a Gaussian random field (GRF) given by

$$p(f|\mathbf{B}) = \text{const} \times |\mathbf{B}|^{1/2} \exp\{-0.5\sigma^{-2}(f - \mu)^T \mathbf{B}(f - \mu)\},$$

where $\mathbf{B} = (b(s_i, s_j))$, the inverse matrix of the covariance matrix of f , controls the spatial dependence structure of f (Besag, 1974). Therefore, we have

$$U(f, \tau) = 0.5\sigma^{-2} f^T \mathbf{B} f - \sigma^{-2} \mu^T \mathbf{B} f + 0.5\sigma^{-2} \mu^T \mathbf{B} \mu,$$

where τ represents all unknown parameters in $(\mathbf{B}, \mu, \sigma)$. In particular, evaluating $|\mathbf{B}|^{1/2}$ is computationally prohibitive, because \mathbf{B} is an $n \times n$ -dimensional matrix (e.g., a 2048×2048 matrix corresponding to a 64×64 lattice) (Rue, 2001; Gu and Zhu, 2001). For edge-preserving image recovery, we further consider a generalized GRF (Bouman and Sauer, 1993) defined as follows:

$$p(f|\mathbf{B}) = \frac{1}{\sigma^N C(\mathbf{B}, p_0)} \exp\left\{-\frac{1}{p_0 \sigma^2} \sum_{s_i \sim s_j} b(s_i, s_j) |f(s_i) - f(s_j)|^{p_0}\right\},$$

where the summation is taken over all nearest-neighbor pairs $(s_i \sim s_j)$, $p_0 \in (1, 2]$, and the normalized constant $C(\mathbf{B}, p_0)$ depends on both $b(s_i, s_j)$ and p_0 .

EXAMPLE 2 (Image segmentation). Image segmentation is used to classify an image into a set of nonoverlapping regions $\{R_1, \dots, R_K\}$. We consider a special case of SMMLs as follows. The observation at a particular pixel s can be written as

$$y(s) = \sum_{k=1}^K \Phi(s, \beta_k) f_k(s) + \varepsilon(s), \quad (4)$$

where $\varepsilon(s) \sim N(0, \phi^{-1})$, $\Phi(\cdot, \cdot)$ is a parametric model, and β_k is the parameter vector for R_k . In addition, $f(s) = (f_1(s), \dots, f_K(s))$, $f_k(s) \in \{0, 1\}$, $\sum_{k=1}^K f_k(s) = 1$, and $f_k(s) = 1$ if and only if $s \in R_k$. Thus, $\mu(s) = \text{E}[y(s)|f] = \sum_{k=1}^K \Phi(s, \beta_k) f_k(s)$. We further assume that the joint distribution of the label field $f = \{f(s) : s = 1, \dots, n\}$ is

given by

$$p(f|\tau) = \exp\left\{\tau \sum_{s_i \sim s_j} \delta(f(s_i), f(s_j)) - \log C(\tau)\right\},$$

where the summation is taken over all nearest-neighbor pairs $(s_i \sim s_j)$, $\delta(x, z)$ is the Kronecker function equaling to 1 when $x = z$ and 0 otherwise, and τ is the parameter controlling the granularity of the field. In addition, $C(\tau)$ is obtained by summing over all possible configurations f (e.g., n^M terms).

3. Estimation procedure

Much effort has been devoted to developing procedures for estimating the parameters and latent fields of SMMLs. See, for example, Marroquin et al. (2003), Lakshmanan and Derin (1989), Jalobeanu et al. (2002), Saquib et al. (1998), Qian and Titterton (1991), Zhu et al. (2005a), and Younes (1989). An approach proposed by Lakshmanan and Derin (1989) is based on jointly maximizing the unknown parameters and the latent fields, but the estimates of parameters under this approach may not be consistent statistically (Neyman and Scott, 1948). For instance, for GLMM, specific conditions are required for validity of this approach (Jiang et al., 2001). To avoid such a pitfall, we take an alternative approach by calculating MLE of $\xi = (\tau, \alpha)$ first and then computing a maximum a posteriori (MAP) estimate of latent field f . In particular, MLE of ξ is a consistent estimate under certain conditions (Guyon, 1995). Thus, our estimation procedure consists of two key steps as follows:

Stage 1: compute MLE of ξ , denoted by $\hat{\xi}$, by using the SAEM algorithm;

Stage 2: given $\hat{\xi}$ obtained in Stage 1, we calculate the MAP estimate of f by using the ASAMC algorithm.

3.1. Stochastic approximation expectation-maximization algorithm

The MLE $\hat{\xi} = (\hat{\tau}, \hat{\alpha})$ is defined by

$$L(\hat{\xi}; y_o) = \max_{\xi} L(\xi; y_o), \quad (5)$$

where y_o denotes the observed data, and the likelihood function of observed-data $L(\xi; y_o)$ is given by

$$L(\xi; y_o) = \int \left[\prod_{i=1}^n p(y(s_i)|f, \mathbf{x}, \alpha) \right] \exp\{-U(f, \tau) - \log C(\tau)\} m(df). \quad (6)$$

The integration above is usually of very high dimension, making direct numerical evaluation difficult even for today's computers. In addition, $C(\tau)$ may involve a large matrix as in Example 1, a huge summation as in Example 2, and so on. In order to obtain $\hat{\xi}$, we assume throughout the paper that $L(\xi; y_o)$ is sufficiently smooth and $\hat{\xi}$ always exists and is unique throughout the paper.

To calculate MLE, we approximate the first-order and second-order derivatives of the log-likelihood function of observed data. From the missing information principle, it follows that the first-order derivative of log-likelihood function can be written as

$$s_{\xi}^k(\xi; y_o) = \partial_{\xi} \log L(\xi; y_o) = E[S_{\xi}^k(\xi; f) | y_o, \xi], \quad (7)$$

where $\partial_{\xi} = \partial/\partial\xi$, $E[\cdot | y_o, \xi]$ denotes the expectation taken with respect to the conditional distribution f given the observed data, and $S_{\xi}^k(\xi; f)$ is the first derivative of complete-data log-likelihood function. Here, the complete-data log-likelihood function $L_c(\xi; f, y_o)$ is given by

$$\sum_{s \in S} \log p(y(s) | f, x, \alpha) - U(f, \tau) - \log C(\tau). \quad (8)$$

To calculate the second-order derivative of the log-likelihood function, we apply Louis's (1982) formula and obtain

$$-\partial_{\xi}^2 \log L(\xi; y_o) = E[I_{\xi\xi}^k(\xi; f) - S_{\xi}^k(\xi; f)^{\otimes 2} | y_o, \xi] + s_{\xi}^k(\xi; y_o)^{\otimes 2}, \quad (9)$$

where for vector \mathbf{a} , $\mathbf{a}\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$, and $I_{\xi\xi}^k(\xi; f) = -\partial_{\xi}^2 L_c(\xi; f, y_o)$ denotes the complete data information matrix.

For SMEMs, we need to approximate the first-order and second-order derivatives of the complex $C(\tau)$. Following Gu and Zhu (2001), we can show that

$$\begin{aligned} \partial_{\tau} \log C(\tau) &= -E_{\tau}[\partial_{\tau} U(f, \tau)], \\ \partial_{\tau}^2 \log C(\tau) &= -E_{\tau}[J(\tau; f)] - \{\partial_{\tau} \log C(\tau)\}^{\otimes 2}, \end{aligned} \quad (10)$$

where $J(\tau; f) = \partial_{\tau}^2 U(f, \tau) - [\partial_{\tau} U(f, \tau)]^{\otimes 2}$ and E_{τ} is taken with respect to MRF (1). Based on (10), we approximate $\partial_{\tau} \log C(\tau)$ and $\partial_{\tau}^2 \log C(\tau)$ by using certain Markov chain Monte Carlo (MCMC) methods, such as the hybrid Markov chain, the birth-and-death process, and the Metropolis-Hastings (MH) algorithm. See, for example, Metropolis et al. (1953), Hastings (1970), Liu (2001), Möller (1999), and Robert and Casella (1999), among others. An alternative approach is to use numerical integration, but it usually gives unstable estimates except in some special cases.

We can approximate the first-order and second-order derivatives of the likelihood functions of observed data by using Eqs. (7), (8), and (10). The $\partial_{\xi} \log L(\xi; y_o)$ can be approximated by $(S_{\tau,1}^T - S_{\tau,2}^T)^T$, $S_{\alpha}(\xi; f)^T$, where $S_{\tau,2} = \partial_{\tau} \log C(\tau)$ and $S_{\tau,1} = -E_{\xi}[\partial_{\tau} U(f, \tau) | y_o]$. We define

$$I_1(\xi; f) = \begin{pmatrix} \partial_{\tau}^2 U(f, \tau) & 0 \\ 0 & I_{\alpha\alpha}(\xi; f) \end{pmatrix}$$

and

$$I_2(\xi; f) = - \begin{pmatrix} -\partial_{\tau} U(f, \tau) \\ S_{\alpha}(\xi; f) \end{pmatrix}^{\otimes 2}.$$

The information matrix $-\partial_{\xi}^2 \log L(\xi; y_o)$ can be approximated by

$$\begin{aligned} E_{\xi}[I_1(\xi; f) | y_o] + \begin{pmatrix} -E_{\tau}[J(\tau; f)] - (S_{\tau,2})^{\otimes 2} & 0 \\ 0 & 0 \end{pmatrix} + E_{\xi}[I_2(\xi; f) | y_o] \\ + \begin{pmatrix} -(S_{\tau,2})^{\otimes 2} + S_{\tau,1} S_{\tau,2}^T + S_{\tau,2} S_{\tau,1}^T & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} + s_{\xi}^k(\xi; y_o)^{\otimes 2}. \end{aligned} \quad (11)$$

3.1.1. Basic steps of the SAEM algorithm

We introduce the SAEM algorithm for SMEMs as follows. We adaptively update seven estimates: $\hat{\xi}^k$, the current estimate of ξ ; $S_{\tau,1}^k$, the current estimate of $E_{\xi}[-\partial_{\tau} U(f, \tau) | y_o]$; $S_{\tau,2}^k$, the current estimate of $-\partial_{\tau} \log C(\hat{\tau})$; \mathbf{h}^k , the current estimate of $s_{\xi}^k(\hat{\xi}; y_o)$; \mathbf{F}_1^k , the current estimate of $E_{\xi}[I_1(\hat{\xi}; f) | y_o]$; \mathbf{F}_2^k , the current estimate of $E_{\xi}[I_2(\hat{\xi}; f) | y_o]$; and \mathbf{F}_3^k , the current estimate of $E_{\xi}[J(\hat{\tau}; f)]$. Let $\Pi_{\tau}(\cdot, \cdot)$ denote the Markov transition probability of the MH algorithm for simulating f from MRF (1), and let $\Pi_{y_o, \xi}(\cdot, \cdot)$ denote the transition probability of the MH algorithm for simulating f conditional on y_o .

Step 1. At the k th iteration, set $f_{k,0} = f_{k-1, N_k-1}$ and $f_{y,k,0} = f_{y,k-1, N_k-1}$. For $i = 1, \dots, N_k$, generate $f_{k,i}$ and $f_{y,k,i}$ from the transition probability $\Pi_{\tau, \xi^{k-1}}(f_{k,i-1}, \cdot)$ and $\Pi_{y_o, \xi^{k-1}}(f_{y,k,i-1}, \cdot)$, respectively.

Step 2. Update the seven estimates as follows:

$$\begin{cases} \hat{\xi}^k = \hat{\xi}^{k-1} + \gamma_k [T_{\xi}^{k-1} \bar{H}(\hat{\xi}^{k-1}; f_k, f_{y,k}), \\ \mathbf{h}^k = \mathbf{h}^{k-1} + \gamma_k (\bar{H}(\hat{\xi}^{k-1}; f_k, f_{y,k}) - \mathbf{h}^{k-1}), \\ \mathbf{F}_1^k = \mathbf{F}_1^{k-1} + \gamma_k (\bar{I}_1(\hat{\xi}^{k-1}; f_{y,k}) - \mathbf{F}_1^{k-1}), \\ \mathbf{F}_2^k = \mathbf{F}_2^{k-1} + \gamma_k (\bar{I}_2(\hat{\xi}^{k-1}; f_{y,k}) - \mathbf{F}_2^{k-1}), \\ \mathbf{F}_3^k = \mathbf{F}_3^{k-1} + \gamma_k (\bar{J}(\tau^{k-1}; f_k) - \mathbf{F}_3^{k-1}), \\ S_{\tau,1}^k = S_{\tau,1}^{k-1} + \gamma_k (-\partial_{\tau} \bar{U}(f_{y,k}, \tau^{k-1}) - S_{\tau,1}^{k-1}), \\ S_{\tau,2}^k = S_{\tau,2}^{k-1} + \gamma_k (\partial_{\tau} \bar{U}(f_k, \tau^{k-1}) - S_{\tau,2}^{k-1}), \end{cases} \quad (12)$$

where $\mathbf{h}^k T = (\mathbf{h}_1^k T, \mathbf{h}_2^k T)$, $f_k = (f_{k,1}, \dots, f_{k, N_k})$ and $f_{y,k} = (f_{y,k,1}, \dots, f_{y,k, N_k})$,

$$\bar{I}_1(\xi; f_{y,k}) = \sum_{i=1}^{N_k} I_1(\xi; f_{y,k,i}) / N_k,$$

$$\bar{I}_2(\xi; f_{y,k}) = \sum_{i=1}^{N_k} I_2(\xi; f_{y,k,i}) / N_k,$$

$$\bar{J}(\tau; f_k) = \sum_{i=1}^{N_k} J(\tau; f_{k,i}) / N_k,$$

$$\mathbf{F}^k = \mathbf{F}_1^k + \mathbf{h}^k T^{\otimes 2} + \begin{pmatrix} -\mathbf{F}_3^k - (S_{\tau,2}^k)^{\otimes 2} \\ 0 \\ 0 \end{pmatrix}.$$

$$\begin{aligned} \partial_\tau \bar{U}(f_{y,k}, \tau) &= \frac{1}{N_k} \sum_{i=1}^{N_k} \partial_\tau U(f_{y,k,i}, \tau), \\ \partial_\tau \hat{U}(f_k, \tau) &= \frac{1}{N_k} \sum_{i=1}^{N_k} \partial_\tau U(f_{k,i}, \tau), \\ \bar{H}(\xi; f_k, f_{y,k}) &= \left([-\partial_\tau \bar{U}(f_{y,k}, \tau) + \partial_\tau \hat{U}(f_k, \tau)]^\top, \frac{1}{N_k} \sum_{i=1}^{N_k} S_k(\xi; f_{y,k,i})^\top \right)^\top. \end{aligned}$$

3.1.2. Gain constants

Gain constants play an essential role in ensuring the convergence of stochastic approximation algorithms. For fixed N_k , the gain constants sequence $\{\gamma_k\}$ must satisfy the following conditions:

$$0 \leq \gamma_k \leq 1 \text{ for all } k, \quad \sum_{k=1}^{\infty} \gamma_k = \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} \gamma_k^2 < \infty. \quad (13)$$

In practice, gain constants are usually defined by $\gamma_k = b_1/(k^{a_1} + b_1 - 1)$, $k = 1, \dots, K_1$, where integer b_1 and real number $a_1 \in (1/2, 1]$ are preassigned and K_1 is determined by some random criteria (Gu and Zhu, 2001; Zhu et al., 2005a). For a given sequence γ_k , the SAEM algorithm iterates Steps 1 and 2 as described above. At the beginning of the SAEM algorithm, we suggest to choosing a small a_1 so that the SAEM algorithm will move quickly towards to the feasible region. When the algorithm starts to stabilize near the neighborhood of MLE, we set a_1 to be close to 1, and a small integer is chosen for b_1 , say, $a_1 = 0.8$ and $b_1 = 2$. At the same time, an averaging procedure is used, with $\tilde{\xi}^0 = \xi^0$, $\tilde{\mathbf{h}}^0 = \mathbf{h}^0$, $\tilde{S}_{\tau,m}^0 = S_{\tau,m}^0$, and $\tilde{\mathbf{F}}_{m'}^0 = \mathbf{F}_{m'}^0$,

$$\tilde{\xi}^k = \tilde{\xi}^{k-1} + (\xi^k - \tilde{\xi}^{k-1})/k, \quad \tilde{\mathbf{h}}^k = \tilde{\mathbf{h}}^{k-1} + (\mathbf{h}^k - \tilde{\mathbf{h}}^{k-1})/k,$$

$$\tilde{S}_{\tau,m}^k = \tilde{S}_{\tau,m}^{k-1} + (S_{\tau,m}^k - \tilde{S}_{\tau,m}^{k-1})/k, \quad \text{and} \quad \tilde{\mathbf{F}}_{m'}^k = \tilde{\mathbf{F}}_{m'}^{k-1} + (\mathbf{F}_{m'}^k - \tilde{\mathbf{F}}_{m'}^{k-1})/k,$$

for $m = 1, 2$ and $m' = 1, 2, 3$. Theoretically, this averaging procedure automatically leads to an optimal convergence without estimating the information matrix (Polyak, 1990; Polyak and Juditski, 1992). Under some conditions, the off-line average $(\tilde{\xi}^{K_1}, \tilde{\mathbf{h}}^{K_1})$ converges to $(\xi, s_\xi(\xi, \mathbf{y}_0))$ almost surely, as $K_1 \rightarrow \infty$ (Zhu et al., 2005a). Finally, we can substitute $\tilde{S}_{\tau,m}^{K_1}$ ($m = 1, 2$) and $\tilde{\mathbf{F}}_{m'}^{K_1}$ ($m' = 1, 2, 3$) into Eq. (11) to estimate $-\partial_\xi^2 \log L(\xi; \mathbf{y}_0)$.

3.2. Annealing stochastic approximation Monte Carlo algorithms

The annealing stochastic approximation Monte Carlo (ASAMC) algorithm originates from the multicanonical algorithm (Berg and Neuhaus, 1991). In the past decade, the multicanonical algorithm has been studied extensively. See, for instance, (1/ k)-ensemble sampling in Hesselbo and Stinchcombe (1995), the Wang-Landau algorithm

in Wang and Landau (2001), the generalized Wang-Landau (GWL) algorithm in Liang (2004, 2005a, 2005b), and the stochastic approximation Monte Carlo (SAMC) in Liang et al. (2005) and Liang (2005c), among others. In particular, the GWL and SAMC algorithms improve the multicanonical algorithm and its variants by introducing the concept of partitioning the sample space and further extending the multicanonical algorithm from discrete system to continuum system. Computational advances, including the GWL and SAMC algorithms, led to possible solutions to many complex statistical problems, such as model selection, highest posterior density region/interval construction, and the Monte Carlo optimization, among others.

3.2.1. Multicanonical algorithm

We use the Ising model as an example to the explicate multicanonical algorithm. The Gibbs distribution of the Ising model on an $L \times L$ lattice space can be written as

$$p(f|\tau) \propto \exp\{-U(f)/\tau\}, \quad f \in S_f, \quad (14)$$

where $U(f) = -\sum_{s_i=s_j} \delta(s_i, s_j)$ and $S_f = \{-1, 1\}^{L^2}$. We are interested in estimating $\Omega(u) = \#\{f: U(f) = u\}$, called the density of states (or spectral density) of the system. We may directly use MCMC algorithms (e.g., the MH algorithm or the Gibbs sampler) to draw samples from $p(f|\tau)$ and then use the simulated samples to estimate $\Omega(u)$. However, conventional MCMC algorithms can become trapped into a local energy minimum indefinitely, rendering the simulation ineffective. The multicanonical algorithm provides an attractive solution to this difficulty. The multicanonical algorithm seeks to draw samples from a modified distribution given by

$$p_m(f) \propto \exp\{-\log \Omega(U(f))\}. \quad (15)$$

If samples can be exactly drawn from (15), then the resulting distribution for U should be uniform distribution, that is, $pu(u) \propto 1$. Thus, the algorithm will not become trapped into a local energy minimum, because sampling from $p_m(f)$ leads to a "free" random walk in the space of energy. However, $\Omega(u)$ is unknown prior to the simulation.

The key idea of the multicanonical algorithm is to iteratively update the approximation of $\Omega(u)$, denoted as $\hat{\Omega}(u)$, then producing Monte Carlo samples from an approximated version of $p_m(f)$. The statistical quantities related to $p(f)$ can then be estimated based on the Monte Carlo samples with the technique of importance sampling. In addition, the multicanonical algorithm is useful for optimization. For instance, a study on protein folding problems (Hansmann and Okamoto, 1997) shows that the multicanonical algorithm is much more efficient than the temperature rescaling-based algorithms, including simulated tempering and simulated annealing (Marinari and Parisi, 1992; Geyer and Thompson, 1995; Kirkpatrick et al., 1983).

3.2.2. Basic steps of the ASAMC algorithm

Let $\tilde{U}(f)$ be the negative of the complete-data log-likelihood function of SMMs, $-\ell_c(\xi; f, \mathbf{y}_0)$. Given ξ in Stage 1, the ASAMC algorithm is then applied to find an optimal configuration of f , denoted by \hat{f} , which minimizes $\tilde{U}(f)$. The ASAMC algorithm comprises four steps as follows:

Step 1. Partition the sample space S_f into M disjoint subregions, E_1, \dots, E_M , and set an arbitrary configuration f_0 , $\mathbf{g}_f^0 = (g_{0,1}, \dots, g_{0,M}) = (0, \dots, 0)$, a pre-specified parameter Δ , $U_{\min}^{(0)}$, and the search space $S_f^{(0)} = \bigcup_{i=1}^M E_i$;

Step 2. At the k th iteration, use the MH algorithm with a global proposal distribution to simulate a sample f_k from the distribution

$$p_{g_k}(f) \propto \sum_{i=1}^{I(U_{\min}^{(k-1)} + \Delta)} \frac{\psi(f)}{e^{g_{k-1,i}}} \delta(f \in E_i),$$

where $\psi(f) = \exp[-\tilde{U}(f)/t_0]$, t_0 is a preassigned number, $\delta(f \in E_i)$ is the Kronecker function equating to 1 when $f \in E_i$ and 0 otherwise, $U_{\min}^{(k-1)}$ is the minimum energy value obtained until the $(k-1)$ th iteration, and $I(z)$ denotes the index of subregion where a sample f with energy $\tilde{U}(f) = z$ belongs to (e.g., $I(\tilde{U}(f)) = j$ for $f \in E_j$);

Step 3. Update the working parameter g_k in the following manner:

$$g_{k,i} = g_{k-1,i} + \gamma_k [\delta(f_k \in E_i) - \pi_i], \quad i = 1, \dots, M,$$

where $\pi_i \in (0, 1)$ and $\sum_{i=1}^M \pi_i = 1$, and nonincreasing sequence γ_k ($k = 1, \dots$) satisfies

$$\gamma_k > 0, \quad \sum_{k=1}^{\infty} \gamma_k = \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} \gamma_k^q < \infty, \quad (16)$$

where $q \in (1, 2)$. Throughout the paper, we set $\gamma_k = [k_0/\max(k_0, k)]^{1/2}$ for some specified value $k_0 > 1$, where $a_2 \in (0.5, 1]$;

Step 4. Increase k to $k+1$ and update $U_{\min}^{(k)}$, M , and the sample space to $S_f^{(k)} = \bigcup_{i=1}^{I(U_{\min}^{(k)} + \Delta)} E_i$.

We have to impose several conditions on the sample space S_f in the ASAMC algorithm. In Step 1, the sample space is usually partitioned into M disjoint subregions as follows: $E_1 = \{f: \tilde{U}(f) \leq u_1\}$, $E_2 = \{f: u_1 < \tilde{U}(f) \leq u_2\}$, \dots , $E_{M-1} = \{f: u_{M-2} < \tilde{U}(f) \leq u_{M-1}\}$, and $E_M = \{f: u_M \geq \tilde{U}(f) > u_{M-1}\}$, where u_1, \dots, u_{M-1} are specified real numbers such that $u_1 < \dots < u_M$. For SMMs, $\psi(f) = \exp[-\tilde{U}(f)/t_0]$ and $w_{\psi,i} = \int_{E_i} \psi(f) m(d f)$ is the partition function of the truncated distribution of f in the subregion E_i . Furthermore, we assume that the sample space S_f is compact. This condition is trivial for some discrete systems, such as the Ising model. However, for continuous systems, we restrict S_f to a set $\{f: \tilde{U}(f) \leq \tilde{U}_{\max}\}$, where \tilde{U}_{\max} is a fixed large value so that the set $\{f: \tilde{U}(f) > \tilde{U}_{\max}\}$ is not of interest.

Two other important features of the ASAMC algorithm are approximately sampling from $p_m(f)$ as in the multicanonical algorithm and updating working estimates g_k s. For simplicity, we temporarily assume that $U_{\min}^{(k-1)}$ is fixed and $I(U_{\min}^{(k-1)} + \Delta) = M$ (Liang et al., 2005). At the k th iteration, we use an MCMC algorithm to draw a sample from the distribution

$$p_{g_k}(f) \propto \sum_{i=1}^M \frac{\psi(f)}{e^{g_{k-1,i}}} \delta(f \in E_i), \quad (17)$$

where $g_k = (g_{k,1}, \dots, g_{k,M}) \in \mathcal{G}$ is an estimate of $(\log w_{\psi,1}, \dots, \log w_{\psi,M})$ until the k th iteration. In practice, for continuum system, we set $\mathcal{G} = [-B, B]^M$ with $B = 10^{100}$. Because adding to or subtracting from g_k a constant will not change $p_{g_k}(f)$, g_k can be kept in the compact set in simulations by adjusting with an additive constant. Under appropriate conditions,

$$g_{k,i} \rightarrow \begin{cases} c + \log(\int_{E_i} \psi(f) m(d f)) - \log(\pi_i + \eta), & E_i \neq \emptyset, \\ -\infty, & E_i = \emptyset, \end{cases} \quad (18)$$

where $\eta = \sum_{j \in \{i: E_j = \emptyset\}} \pi_j / (M - M_0)$ and M_0 is the number of empty subregions, as $k \rightarrow \infty$ (Liang et al., 2005). In addition, c is a constant which can be determined by imposing a constraint on g_k . For instance, $\sum_{i=1}^M e^{g_{k,i}}$ is equal to a fixed number. Since the sample space is partitioned blindly in the ASAMC algorithm, some of the subregions may be empty, that is, $\int_{E_i} \psi(f) d f = 0$. The working distribution $p_{g_k}(f)$ is obtained by a piecewise modification of $p(f|\tau)$, where each subregion is associated with a different weight $e^{g_{k,i}}$ (Liang et al., 2005).

The above stochastic approximation algorithm is an annealing algorithm, because the sample space $S_f^{(k)}$ shrinks during each iteration. Theoretically, the ASAMC algorithm can find the global energy minimum if the algorithm is run long enough, but the process of locating the global energy minimum may be very slow due to the breadth of the sample space. To accelerate the process, Liang (2004, 2005c) proposed to restrict the sample space of the ASAMC algorithm to a small region during each iteration. Suppose that the subregions E_1, \dots, E_M have been arranged in ascending order by energy; that is, if $i < j$, then $U(f) < U(f')$ for any $f \in E_i$ and $f' \in E_j$. The ASAMC algorithm starts with $S_f^{(0)} = \bigcup_{i=1}^M E_i$, and then iteratively sets

$$S_f^{(l)} = \bigcup_{i=1}^{I(U_{\min}^{(l-1)} + \Delta)} E_i. \quad (19)$$

Remarkably, the ASAMC algorithm preserves the convergence (18) on the limiting sample space $\lim_{l \rightarrow \infty} S_f^{(l)}$, provided that the proposal distribution used at each iteration is global, that is, a proposal distribution $q(f, f')$ is global if $q(f, f') > 0$ for all $f, f' \in S_f$.

As known by many researchers, the state of the art algorithm for stochastic optimization is the simulated annealing algorithm (Kirkpatrick et al., 1983). For instance, we consider the problem of minimizing the function $\tilde{U}(f)$. Simulated annealing works by simulating from a sequence of distributions scaled by the temperature as follows,

$$p_{t_k}(f) \propto \exp\{-\tilde{U}(f)/t_k\}, \quad k = 1, 2, \dots,$$

where t_k 's are called the temperatures forming a decreasing ladder $t_1 > \dots > t_k > \dots \geq 0$. Under some conditions, simulated annealing will converge to the set of global minima of $\tilde{U}(f)$ in probability 1 when the temperature decreases sufficiently slowly, i.e., $t_k > 1/\log(L_k)$, where $L_k = N_1 + \dots + N_k$ (Geman and Geman, 1984). In addition, N_k is the number of iterations generated from the MH algorithm in simulating from the distribution $p_{t_k}(f)$. In practice, such a slow cooling scheme is impractical.

Instead, people use a linearly or geometrically decreasing cooling scheme, but such scheme cannot guarantee that the global minima will be reached. The ASAMC algorithm does not suffer from such a pitfall. If the proposal distribution is global and the gain constants satisfy the condition (16), the ASAMC algorithm will converge to the set of global minima as the number of iterations is large. The ASAMC algorithm will result in a "free" random walk in the subspace of the subregions. Its self-adjusting ability for the acceptance of a new proposal guarantees that it will not become stuck into a local energy minimum. Hence, as a stochastic optimization algorithm, the ASAMC algorithm is potentially much more powerful than the simulated annealing algorithm (Liang, 2005c).

3.2.3. Practical issues

For an effective implementation of the ASAMC algorithm, several issues need to be considered.

(i) *Partitioning the sample space*: For optimization problems, the partition can be done according to the energy function. The maximum energy difference in each subregion should be bounded by a reasonable number, say, 2, to ensure that a reasonable acceptance rate is achieved for the local MH moves within the same subregion. Note that within the same subregion, sampling from the working density (17) reduces to sampling from $\psi(f)$.

(ii) *Choice of Δ* : The performance of the ASAMC algorithm depends on the value of Δ to some extent. If Δ is too large, the ASAMC algorithm may take a long time to locate the global minimum due to the breadth of the sample space. If Δ is too small, the ASAMC algorithm may also take a long time to locate the global minimum. In this case, the sample space may contain only a few isolated regions, and most of the proposed transitions will be rejected. Allowing a sampler to jump to intermediate states of high energy will increase the probability of transition from one local energy minimum to others. To compensate for the negative effect of the sample space restriction, the proposal distribution used in the ASAMC algorithm should be spread out.

(iii) *Choice of k_0 and the number of iterations*: The γ_k controls the moving ability of the ASAMC algorithm across subregions, and k_0 controls the speed of γ_k converging to zero. In practice, k_0 can be chosen according to the complexity of the problem. The more complex the problem, the larger value of k_0 . A large value of k_0 will force the sampler to reach all subregions quickly, even in the presence of multiple local energy minima. The appropriateness of the choice of k_0 and the number of iterations can be diagnosed by examining the convergence of the run, which can be further diagnosed by examining the equality of the realized sampling frequencies of limiting subregions. As suggested by Wang and Landau (2001), a run can be regarded as converged if the sampling frequency for each of the subregions is not less than 80% of the average sampling frequency; that is,

$$\min \left\{ \frac{e_i}{\bar{e}} : i = 1, \dots, l(U_{\min} + \Delta), E_i \neq \emptyset \right\} \geq 80\%, \quad (20)$$

where e_i denotes the realized sampling frequency of the subregion E_i , and \bar{e} is the average sampling frequency of the subregions included in the above set. If a run does

not converge, the ASAMC algorithm should be re-run with more iterations or a larger value of k_0 .

(iv) *Choosing the proposal function*: In the ASAMC algorithm, the global proposal distribution ensures the ergodicity of the algorithm. In practice, a global proposal distribution can be designed easily for both discrete and continuum systems. For example, in simulations from an Ising model of linear size L , a new configuration can be generated with the following steps: draw an integer T with probability e_t ($t = 1, \dots, L^2$), $0 < e_t < 1$ and $\sum_{t=1}^{L^2} e_t = 1$; choose T spins from the set $S = \{(i, j) : i, j = 1, \dots, L\}$ at random and with replacement; reset the value of each of the T spins to $+1$ or -1 with equal probability. We will call this Sampling Method (I). For a particular configuration generated with the above procedure, the transition probability is then $q(f, f') = e_T / 2^T$. A typical choice for the e_t 's is $e_t = 0.9$ and $e_t = (1 - e_t) / (L^2 - 1)$ for $t = 2, \dots, L^2$. For a continuum system, $q(f, f')$ can be set to the random walk Gaussian proposal $f' \sim N(f, \sigma^2)$, with σ^2 being calibrated to have a desired acceptance rate, such as 0.25.

4. Applications

In this section, we analyzed two real data sets from imaging studies from ecology. They will be discussed to illustrate the behavior of the SAEM algorithm, the ASAMC algorithm, and their combination. All computations were done in C++ on a Dell laptop. All computer codes and executable files can be downloaded from Dr. Zhu's website:

<http://www.bios.unc.edu/~hzhn/SMM/smm.tar>.

4.1. Distributions of vegetation species

We consider an automulticategorical model to analyze the dataset of vegetation species in Alberta, Canada. The primary goal of this data analysis is to demonstrate the efficiency of the stochastic approximation algorithm in locating MLE in complex spatial models. In particular, through this example, we want to show the feasibility of roughly approximating the first-order and second-order derivatives of the partition function during each iteration and controlling amount of noises by using stochastic approximation (Robbins and Monro, 1951). The secondary goal is to illustrate the wide application of spatial models.

Vegetation species is in the form of an atlas map with resolution pixel equating 0.5°C latitude $\times 0.5^\circ \text{C}$ longitude; see Figure 1(a). With the aid of remote sensing and aerial photogrammetric technologies, information on four species occurrence in Alberta, Canada is documented by this format (Little, 1971; Arnold, 1993, 1995; Mitchell-Jones et al., 1999). There are total of 375 grid cells. At each site (k, l) , there are a categorical response $Y(k, l)$ and 2 interesting climate covariates: $X_1(k, l)$ (absolute minimum temperature); and $X_2(k, l)$ (annual degree-days). Five major types of vegetation in Alberta are: V0 – Background, V1 – subarctic evergreen forest, V2 – boreal evergreen forest, V3 – boreal summergreen woodland, and V4 – grass prairie. Two covariates are expected to

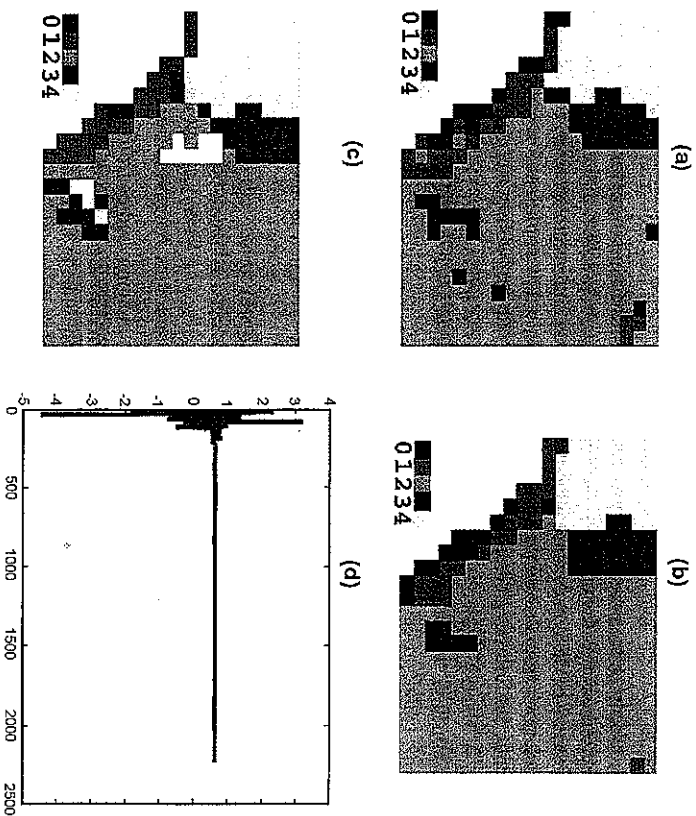


Fig. 1. Distribution of four vegetation types in Alberta, Canada: 0 = background, 1 = subarctic evergreen forest, 2 = boreal evergreen forest, 3 = boreal summergreen woodland, and 4 = grass prairie. There are four panels: (a) the observed distribution; (b) the fitted distribution; (c) the predicted distribution with annual degree-days ($X_2(k, l)$) being increased 350; and (d) $\tau(1)^k$ at each iteration of the SAEM algorithm.

be among those determining the distributions of vegetation at geographical scales and having significant changes in global warming.

Following Zhu et al. (2005b), the second-order automulticategorical regression model is assumed for $Y = \{Y(k, l), (k, l) \in S\}$, where the conditional probability at site $(k, l) \in S$ given all the other values $Y(m, n)$ ($(m, n) \neq (k, l)$) is given as follows

$$\Pr\{Y_{k,l} = i \mid \text{all other sites}\} = \frac{\exp\{g_{k,l}(i|\theta)\}}{\sum_{j=0}^4 \exp\{g_{k,l}(j|\theta)\}}, \quad i = 0, \dots, 4. \quad (21)$$

In addition, $g_{k,l}(i|\theta) = X(k, l)^T \beta(i) + \tau(i) y_{k,l}^*(i)$ for $i = 0, \dots, 4$, where $y_{k,l}^*(i)$ is the number of eight sites $\{(k, l-1), (k, l+1), (k-1, l), (k+1, l), (k-1, l-1), (k+1, l+1), (k-1, l+1), (k+1, l-1)\}$ colored i . To avoid redundancy, we assume that $\beta(0) = \mathbf{0}$ and $\tau(0) = 0$. The SAEM algorithm with $(a_1, b_1) = (0.8, 4)$ and $N_k = 5000$ was used to find the maximum likelihood estimates. The algorithm converged in 2231 iterations. The initial value for the vector ξ was set to be $\xi^0 = \mathbf{0}$.

Table 1
Maximum likelihood estimates of the automulticategorical model to the distribution of vegetation species data in Alberta, Canada

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	EST	SD	EST	SD	EST	SD	EST	SD
$\beta(1)$			$\beta(2)$		$\beta(3)$		$\beta(4)$	
Intercept	2.021	8.765	-5.269	9.003	-7.887	9.993	-18.112	15.700
$X_1(k, l)$	0.129	0.184	0.041	0.188	0.151	0.201	0.428	0.309
$X_2(k, l)$	0.003*	0.001	0.005*	0.001	0.009*	0.002	0.018*	0.005
$\tau(1)$			$\tau(2)$		$\tau(3)$		$\tau(4)$	
	0.624*	0.161	0.526*	0.111	0.383*	0.138	0.779*	0.308

* represents that parameters are different from zero at the significance level $\alpha = 0.05$.

The obtained results are summarized in Table 1. The distribution of the vegetation species is related to the annual degree-days. The autocorrelation coefficient $\tau(i)$ ($i = 1, 2, 3, 4$) are significantly different from zero. The fitted map Y of the distribution of four vegetation types in Alberta is shown in Figure 1(b). Figure 1(d) shows the estimate $\tau(1)^k$ at each iteration of the stochastic approximation algorithm. We observe that our stochastic approximation algorithm is robust to the initial value of ξ and can find MLE.

An important scientific issue is to predict the redistribution of vegetation species under various global warming scenarios. For instance, an enhanced greenhouse effect (e.g. the doubling of atmospheric concentration of CO_2) would increase the global mean temperature from 1.5 to 4.5°C in the future 30 to 50 years. One advantage of the automulticategorical model is that the climate change effect can be quantified through the odds ratio of the conditional probabilities. For example, if the annual degree-days X_2 increases by 350 (approximately equivalent to 1°C increase in daily temperature) while other variables remain constant, then the odds ratio of the conditional probabilities of the j th vegetation presence is supposed to increase by a factor $e^{350\beta_j(i)}$. This result suggests that subarctic evergreen forest, boreal evergreen forest, boreal summergreen woodland, and grass prairie will be increased by global warming. The impact of climate change on the distributions of four vegetation types is shown in Figure 1(c).

4.2. Simulation study

Consider a degraded pixel image on a finite grid S of pixels, placing a binary random variable $Y(i, j)$ at each site (i, j) on S , a subset of a regular $M_0 \times N_0$ lattice. Let the true image be $f = \{f(k, l): (k, l) \in S\}$, where $f(k, l) = 0$ represents a white pixel and $f(k, l) = 1$ represents a black pixel. Because S is usually a irregular lattice in most applications, we consider the joint distribution of the internal site responses $f^I = \{f(k, l): (k, l) \in S^o\}$ conditional upon fixed boundary values $f^B = \{f(k, l): (k, l) \in \partial S\}$, where ∂S and S^o denote the set of all sites forming the boundary of S and the set of all internal sites of S , respectively. Following Besag (1974), the probability function

of the first-order autologistic regression of f^l given f^b can be written in a Gibbsian form as follows:

$$p(f^l | \tau, f^b) = \exp\{\tau^T T(f)\} / C(\tau), \quad (22)$$

where $T(f) = \sum_{(k,l) \in S^o} f(k,l) \tilde{X} f(k,l)$ and $\tilde{X} f(k,l) = (X(k,l)^T, \tilde{f}(k,l)/2)^T$, in which $\tilde{f}(k,l)$ is the number of sites in $\{(k,l-1), (k,l+1), (k-1,l), (k+1,l)\}$ colored 1. Also, $X(k,l)$ is a $p \times 1$ vector of covariates at site (k,l) , $\tau = (\beta^T, \tau_1)^T \in R^{p+1}$, $\beta \in R^p$, and $\tau_1 \in R$. Given the true image f , the true observed image $Y = \{Y(k,l) : (k,l) \in S\}$ is assumed to be conditionally mutually independent and

$$Y(k,l) | f(k,l) \sim \text{Binomial}(1, p(f(k,l))), \quad (23)$$

where $p(0) = 0$ and $p(1) = \exp(-\alpha(1)^2) \in (0, 1]$. That is, if $f(k,l) = 0$, $Y(k,l) = 0$ with probability 1, while $Y(k,l) = 1$ with probability $p(1)$ and $Y(k,l) = 0$ with probability $1 - p(1)$ under $f(k,l) = 1$. Thus, we obtain an SMM. In practice, scientists may consider the first-order, second-order and even higher correlation structures; see, for example, the first-order structure in Huffer and Wu (1998) and the second-order structure in Besag (1974, 1986) and He et al. (2003).

In order to check the usefulness of the proposed algorithm, we consider the following simulation study, in which the autologistic regression model is set on a 30×30 lattice and $X(k,l) = (2.5 \times \sin(0.1 \times (k+l)))$. In our simulation, $\beta = 1$, $\tau_1 \in \{0.2, 0.4, 0.6, 0.8\}$, and $\alpha(1) = 0.85(p(1) \approx 0.325)$. Therefore, there are three unknown parameters. To simulate the process $f = \{f(k,l) : (k,l) \in S^o\}$ from (22), we use the standard Gibbs sampler. The initial state of the process is taken at random such that $X(k,l)$ is independently taken to be 1 or 0 with 1/2 probability and the Gibbs sampler is repeated 10 000 times (10 000 Monte Carlo steps) to ensure that the equilibrium state is achieved. Afterwards, the binomial noise is added according to (23).

For each parameter vector $\xi = (\tau_1, \beta, p(1))^T$, we generated $N = 500$ datasets. For each pseudo-observed dataset, the SAEM algorithm with $(a_1, b_1) = (0.8, 5)$ was applied to get the MLE of the unknown parameters. The initial value of ξ was set at $(0, 0, 0.5)$. In each iteration of the algorithm, the standard Gibbs sampler was used to generate f from $p(f|\tau)$; therefore, we can estimate $q, \log C(\tau)$ and $q^2 \log C(\tau)$. To simulate the process Y given f , we used the following algorithm. If $f(k,l) = 1$, $Y(k,l)$ must be equal to 1; however, given $Y(k,l) = 0$ and other $f(u,v)$ s, we have

$$P(f(k,l) = 1 | Y(k,l) = 0, \text{all other values}) \\ = \frac{(1 - p(1)) \exp(X(k,l)^T \beta + \tilde{f}(k,l) \tau_1)}{1 + (1 - p(1)) \exp(X(k,l)^T \beta + \tilde{f}(k,l) \tau_1)}.$$

The standard Gibbs sampler is also used. The number N_k was set at 30.

In this simulation study, $\beta = 1$ represents relatively strong covariate effect and τ_1 ranges in four different cases. We calculated the bias, the mean of the standard deviation estimates, and the root mean-square error obtained from the 500 estimates. The results obtained are summarized in Table 2. It can be seen that all the relative efficiencies are close to 1.0. This demonstrates that the SAEM algorithm is a useful method for optimizing SMMs.

Table 2
Bias, RMS, SD, and EFF of the maximum likelihood estimators of the noisy autologistic regression model

True	Bias	RMS	SD	EFF	True	Bias	RMS	SD	EFF		
β	1.0	-0.008	0.081	0.081	1.005	β	1.0	-0.004	0.080	0.079	1.017
τ_1	0.2	0.003	0.082	0.081	1.007	τ_1	0.4	0.004	0.073	0.077	0.945
$p(1)$	0.325	0.009	0.064	0.064	0.997	$p(1)$	0.325	0.007	0.043	0.044	0.978
β	1.0	-0.010	0.081	0.076	1.066	β	1.0	-0.002	0.083	0.083	0.991
τ_1	0.6	-0.027	0.075	0.074	1.009	τ_1	0.8	0.002	0.083	0.087	0.955
$p(1)$	0.325	0.006	0.033	0.034	0.962	$p(1)$	0.325	0.003	0.028	0.027	1.011

True denotes the true value of parameters; Bias denotes the bias of the mean of estimates; RMS denotes the root-mean-square error; SD denotes the mean of standard deviation estimates; and EFF denotes the ratio of SD and RMS.

4.3. Noisy vegetation data

We analyzed a real data on the distribution of subarctic evergreen woodland vegetation in terms of climate variables in the province of British Columbia, Canada. The subarctic evergreen woodland is in the form of an atlas map with resolution pixel equalling 0.5°C latitude $\times 0.5^\circ\text{C}$ longitude. The observed map Y of the subarctic evergreen woodland is shown in Figure 2(a). There are total of 707 grid cells. At each site (k,l) , there are a binary $Y(k,l)$ and 5 climate covariates of interest: $X_1(k,l)$ (absolute minimum temperature); $X_2(k,l)$ (annual degree-days); $X_3(k,l)$ (total actual evapotranspiration); $X_4(k,l)$ (annual soil moisture deficit); and $X_5(k,l)$ (annual snowpack). These covariates are expected to be among those determining the distribution of vegetation at geographical scales, and they are also the variables likely to change significantly because of global warming. Here, $Y(k,l) = 1$ indicates that at least one subarctic evergreen woodland vegetation has been observed and $Y(k,l) = 0$ indicates that subarctic evergreen woodland are either not inhabited or the subarctic evergreen woodland have not been observed. It is an obvious idea to interpret the observed map in Figure 2(a) as a degraded pixel image, in which a part of the originally black squares are white in the observed map.

We fitted the dataset by the noisy autologistic regression model (22) and (23). The stochastic approximation algorithm with $(a_1, b_1) = (0.8, 4)$ and $N_k = 30$ was used to find the MLE. The initial value of $\xi = (\tau_1, \beta, p(1))$ was set to be $\xi^0 = (0, 0, 0.5)$. The obtained results are summarized in Table 3. The distribution of subarctic evergreen woodland vegetation is related to the absolute minimum temperature and the annual degree-days. The autocorrelation coefficient τ_1 is as high as a value at 1.52. Figure 2(c) shows the trace of τ_1^k , $\tilde{\tau}_1^k$ and $(p(1))^k$, $\tilde{p}(1)^k$ at each iterations of the SAEM algorithm. We compared ICM with the simulated annealing and ASAMC algorithms, which search for the MAP \hat{f} by minimizing $\tilde{U}(f) = -\ell(f|Y; \hat{\xi})$ (except for a constant) given by

$$-\tilde{\tau}^T T(f) - \sum_{(k,l) \in S^o} \log[(1 - \hat{p}(f(k,l)))^{1-Y(k,l)} \hat{p}(f(k,l))^{Y(k,l)}],$$

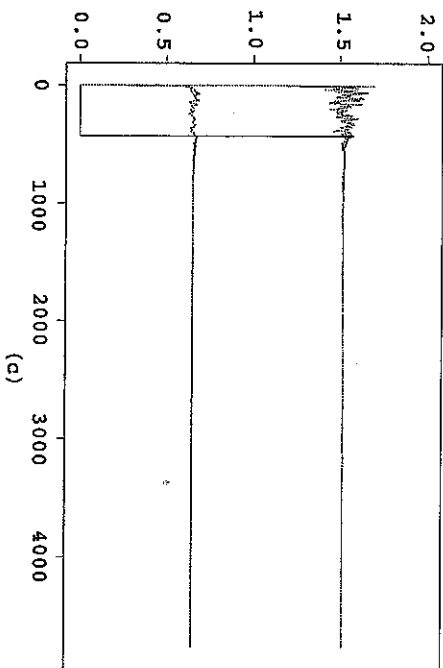
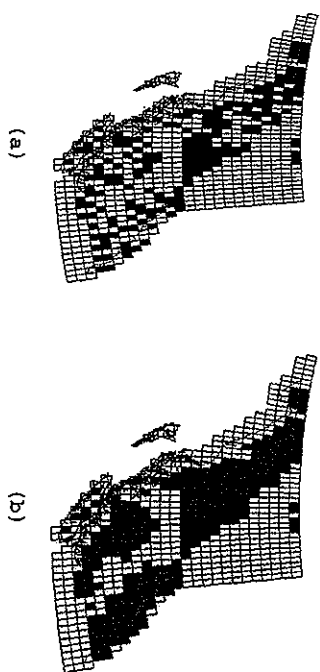


Fig. 2. Subarctic evergreen woodland data: (a) observed map; (b) restored map by the ASAMC algorithm; (c) $(\tau_1^k, \tilde{\tau}_1^k)$ and $(\alpha(1)^k, \tilde{\alpha}(1)^k)$ at each iteration of the SAEM algorithm.

where $T(f) = \sum_{(k,l) \in S^0} f(k,l) \tilde{X}(k,l)$. The $\tilde{U}(f)$ contains $N_0 = 496$ variables $\{f(k,l): Y(k,l) = 0\}$, because if $Y(k,l) = 1$, $f(k,l)$ must be one. Thus, the sample space has $2^{N_0} = 2^{496}$ configurations and direct searching for global minima is therefore nearly infeasible computationally. The ICM method (Besag, 1986) converged in only three iterations and leads to a local minimum 322.935. The simulated annealing and the ASAMC algorithms were run for 2×10^6 iterations. The simulated annealing located a local minimum close to 312.638, while the ASAMC algorithm located an energy minimum at 311.8. This demonstrates that the simulated annealing and ASAMC algorithms require much more computational cost, but they are able to locate the global energy minima with high probability. We include MAP \hat{f} estimated from the ASAMC algorithm in Figure 2(b).

For the simulated annealing algorithm, we considered a linear cooling scheme. We set the highest temperature $T_1 = 10$, the total number of temperature levels $K = 400$,

Table 3
Model fits to the subarctic evergreen woodland data

Iteration	MCMC-SA algorithm ($\alpha = 0$)					τ_1	$p(1)$	
	const	X_1	X_2	X_3	X_4			X_5
5391 EST	-0.7608	0.0371*	-0.0012*	0.0091*	0.0024	-0.0004	1.5190*	0.6534*
650s SD	0.9639	0.0154	0.0004	0.0046	0.0016	0.0003	0.1493	0.0695

* represents that parameters are different from zero at the significance level $\alpha = 0.05$.

and the lowest temperature $T_{400} = 0.01$. At the T_i temperature level, we set $\varepsilon_1 = 0.5 + (10 - T_i)/25$ and $\varepsilon_l = (1.0 - \varepsilon_1)/(N_0 - 1)$ in Sampling Method (I) and run this procedure for 5000 iterations. The temperature decreased linearly such that $T_i = T_{i-1} - \rho$, where $\rho = (T_1 - T_{400})/(K - 1) \approx 2.504 \times 10^{-2}$. The initial configuration of $\{f(k,l): Y(k,l) = 0\}$ was set as $\{f(k,l) = 0: Y(k,l) = 0\}$.

We presented in Figure 3(a) the index plot of the minimum values of $\tilde{U}(f)$ at each temperature level and included in Figure 3(b) the index plots of the values of $\tilde{U}(f)$ in the first and last 15 000 iterations. We observed that the simulated annealing algorithm led to a random walk in the sample space at high temperature. However, the simulated annealing algorithm became trapped in a local minimum 312.877 at low temperature, even though it located a minimum value at 312.638 across all temperature levels.

We applied the ASAMC algorithm to search for the minimum energy value of $\tilde{U}(f)$ by using the following settings. The sample space was partitioned into $M = 1998$ subregions with an equal energy bandwidth: $E_1 = \{f: \tilde{U}(f) \leq 310.599\}$, $E_2 = \{f: 310.599 < \tilde{U}(f) \leq 310.849\}$, ..., and $E_{1998} = \{f: \tilde{U}(f) \leq 810.059\}$. We set $\psi(f) = \exp(-\tilde{U}(f)/10)$, $\pi_1 = \dots = \pi_M = 1/M$, $\tilde{U}_{\max} = 810.059$, $a_2 = 0.6$, $k_0 = 2500$, and $\Delta = 5$. The total number of iterations of the ASAMC algorithm was set at 2×10^6 , which is the same as that of the simulated annealing algorithm. We chose the proposal distribution of Sampling Method (I), in which $\varepsilon_l = (1 - \varepsilon_1)/(N_0 - 1)$ for $l = 2, \dots, N_0$ and ε_1 was set as 0.9 for $0 \leq k \leq 5 \times 10^5$, 0.7 for $5 \times 10^5 < k \leq 10^6$, and 0.5 for $k > 10^6$. The initial configuration of $\{f(k,l): Y(k,l) = 0\}$ was set as $\{f(k,l) = 0: Y(k,l) = 0\}$.

The ASAMC algorithm outperforms the simulated annealing algorithm in this complex noisy vegetation data. In Figure 3(c), we presented the index plots of the values of $\tilde{U}(f)$ at the 5000kth iteration and the minimum values of $\tilde{U}(f)$ until the 5000kth iteration from the ASAMC algorithm, where $k = 0, \dots, 400$. We observed that the ASAMC algorithm converges very quickly to a global minimum of $\tilde{U}(f)$ at 311.8; in contrast, the simulated annealing algorithm wandered around in the sample space at the high temperature and became trapped in local minima at the low temperature (Figure 3(b)). Figure 3(d) shows the sampling frequency of each of the subregions of the ASAMC run. The 10 subregions with the lowest $\tilde{U}(f)$ values are sampled approximately evenly. This indicates that the run has converged. Recall the diagnostic criterion given in (20) for the convergence of the ASAMC runs.

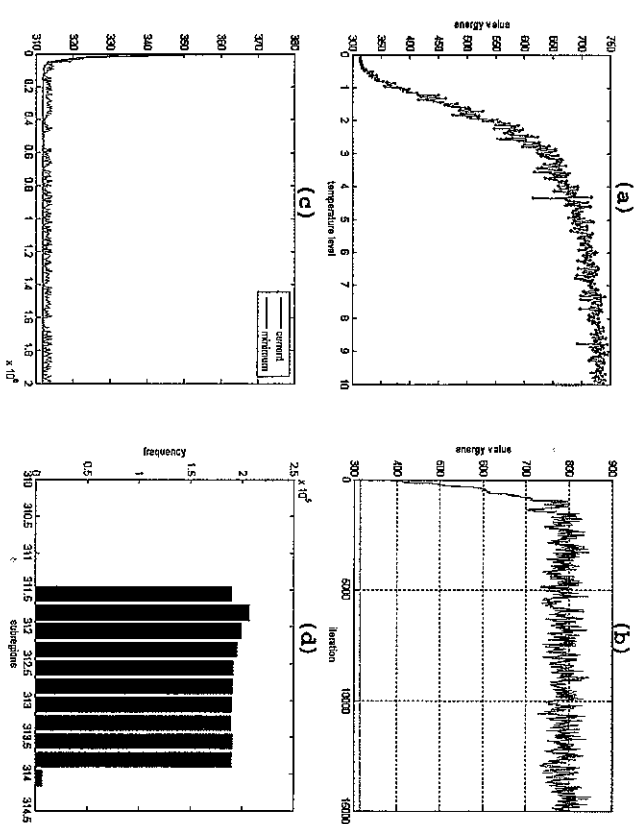


Fig. 3. Subarctic evergreen woodland data: comparison of the simulated annealing and ASAMC algorithms. $\tilde{U}(f)$ is the negative value of the log-likelihood function of complete data, $-\ell(\hat{\xi}, f; y_0)$. There are four panels: (a) the index plot of the minimum energy values of $\tilde{U}(f)$ at each temperature level from the simulated annealing algorithm; (b) the index plots of the energy values of $\tilde{U}(f)$ in the first (red line) and last (green line) 15 000 iterations from the simulated annealing algorithm; (c) the index plots of the energy values of $\tilde{U}(f)$ (red line) at each of the 5000th iterations and the minimum energy values of $\tilde{U}(f)$ (blue line) until the 5000th iteration from the ASAMC algorithm, where $k = 0, \dots, 400$; (d) the sampling frequency in last ten subregions from the ASAMC algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this chapter.)

Acknowledgements

This work was supported in part by NSF grant SES-0643663 to Dr. Zhu, NSF grant DMS-0405748 and NCI grant CA104620 to Dr. Liang, by NIDA grant DA017820, NIMH grants MH068318 and KO274677 to Dr. Peterson, by the Suzanne Crosby Murphy Endowment at Columbia University College of Physicians and Surgeons, and by the Thomas D. Klingenstein and Nancy D. Perlman Family Fund. Thanks to Dr. Jason Royal for his invaluable editorial assistance and to Dr. Fangliang He for making his vegetation dataset available to us.

References

Arnold, H.R. (1993). *Atlas of Mammals in Britain*. Her Majesty's Stationery Office, London.
 Arnold, H.R. (1995). *Atlas of Amphibians and Reptiles in Britain*. Her Majesty's Stationery Office, London.

- Bentler, P.M., Dudgeon, P. (1996). Covariance structure analysis: statistical practice, theory, and directions. *Annals of Review Psychology* **47**, 563–592.
- Berg, B.A., Neuhaus, T. (1991). Multicategorical algorithms for 1st order phase-transitions. *Physics Letters B* **267**, 249–253.
- 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.E. (1986). On the statistical analysis of dither pictures (with discussion). *Journal of the Royal Statistical Society, Series B* **48**, 259–302.
- Booth, J.G., Hobert, J.P. (1999). Maximum generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* **61**, 265–285.
- Bouman, C., Sauer, K. (1993). A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Transactions in Image Processing* **2**, 296–310.
- Breslow, N.E., Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.
- Delyon, B., Lavallée, E., Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* **27**, 94–128.
- Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- Digggle, P.J., Tawn, J.A., Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299–350.
- Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
- Geyer, C.J., Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society, Series B* **54**, 657–699.
- Geyer, C.J., Thompson, E.A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90**, 909–920.
- Gu, M.G., Kong, F.H. (1998). A stochastic approximation algorithm with Markov chain Monte Carlo method for incomplete data estimation problems. *Proceeding of National Academic Science of USA* **95**, 7720–7724.
- Gu, M.G., Zhu, H.T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society, Series B* **63**, 339–355.
- Guyon, X. (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer-Verlag, Berlin.
- Hausmann, U.H.E., Okamoto, Y. (1997). Numerical comparisons of three recently proposed algorithms in the protein folding problems. *Journal of Computational Chemistry* **18**, 920–933.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57**, 97–109.
- He, F.L., Zhou, J.L., Zhu, H.T. (2003). Autologistic regression model for the distribution of vegetation. *Journal of Agricultural, Biological and Environmental Statistics* **8**, 205–222.
- Hesselbo, B., Stinchcombe, R.B. (1995). Monte Carlo simulation and global optimization without parameters. *Physical Review Letters* **74**, 2151–2155.
- Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.
- Huffer, F.W., Wu, H.L. (1998). Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. *Biometrics* **54**, 509–524.
- Jalobeanu, A., Blanc-Feraud, L., Zerubia, J. (2002). Hyperparameter estimation for satellite image restoration using a MCMC maximum-likelihood method. *Pattern Recognition* **35**, 341–352.
- Jiang, J.M., Jia, H.M., Chen, H. (2001). Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica* **11**, 97–120.
- Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.
- Lakshmanan, S., Derrin, H. (1989). Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**, 954–963.

- Lee, S.Y., Zhu, H.T. (2000). Statistical Analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology* 53, 209–232.
- Lee, S.Y., Zhu, H.T. (2002). Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika* 67, 189–210.
- Li, S.Z. (2001). *Markov Random Field Modeling in Image Analysis*. Springer-Verlag, Tokyo.
- Liang, F. (2004). Annealing contour Monte Carlo for structure optimization in an off-lattice protein model. *Journal of Chemical Physics* 120, 6756–6763.
- Liang, F. (2005a). A generalized Wang-Landau algorithm for Monte Carlo computation. *Journal of the American Statistical Association* 100, 1311–1327.
- Liang, F. (2005b). Evidence evaluation for Bayesian neural networks. *Neural Computation* 17, 1385–1410.
- Liang, F. (2005c). Annealing stochastic approximation Monte Carlo for neural network training. *Machine Learning* (revised).
- Liang, F., Liu, C., Carroll, R.J. (2005). Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association*, in press.
- Little Jr., E.J. (1971). *Atlas of United States Trees*, vol. 15. U.S. Government Printing Office, Washington, DC.
- Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.
- Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* 44, 190–200.
- Marnani, F., Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters* 19, 451–458.
- Marroquin, J.L., Santana, E.A., Boleto, S. (2003). Hidden Markov measure field models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 1380–1387.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21, 1087–1091.
- McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 92, 162–170.
- Mitchell-Jones, A.J., Amori, G., Bogdanowicz, W., Kryszufek, B., Reijnders, P.J.H., Spitznberger, F., Stubbe, M., Thissen, J.B.M., Vohralik, V., Zima, J. (1999). *The Atlas of European Mammals*. Poyser, London.
- Müller, J. (1999). Markov chain Monte Carlo and spatial point processes. In: Kendall, W.S., Barndorff-Nielsen, O.E., van Lieshout, M.C. (Eds.), *Stochastic Geometry: Likelihood and Computation*. Chapman and Hall, London.
- Neyman, J., Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.
- Polyak, B.T. (1990). New stochastic approximation type procedures. *Automata i Telemekh.* 98–107. English transl. in: *Automat. Remote Control* 51.
- Polyak, B.T., Juditski, A.B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal of Control and Optimization* 30, 838–855.
- Qian, W., Titterton, D.M. (1991). Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London, Ser. A* 337, 407–428.
- Robbins, H., Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics* 22, 400–407.
- Robert, C.P., Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.
- Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B* 63, 325–338.
- Saguth, S.S., Bouman, C.A., Sauer, K. (1998). ML parameter estimation for Markov random fields with applications to Bayesian tomography. *IEEE Transactions on Image Processing* 7, 1029–1044.
- Vasconcelos, N., Lippman, A. (2001). Empirical Bayesian motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 217–221.
- Wang, F., 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., Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor man's data augmentation algorithm. *Journal of the American Statistical Association* 85, 699–704.

- Winkler, G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods: A Mathematical Introduction*. Springer-Verlag, Berlin, Heidelberg.
- Younis, L. (1989). Parameter estimation for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields* 82, 625–645.
- Zeger, S.L., Liang, K.Y., Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44, 1049–1060.
- Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics* 56, 129–136.
- Zhang, H.P. (1993). Image restoration: flexible neighborhood systems and iterated conditional expectations. *Statistica Sinica* 3, 117–139.
- Zhu, H.T., Lee, S.Y. (2002). Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte Carlo method. *Statistics and Computing* 12, 175–183.
- Zhu, H.T., Gu, M.G., Peterson, B.S. (2005a). Maximum likelihood from spatial random effect model via the stochastic approximation EM algorithm. *Statistics and Computing*, in press.
- Zhu, H.T., He, F.L., Zhou, L.J. (2005b). A nonmulticategorical regression model for the distributions of vegetation types. *Biometrics* (revised).