



## Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications

David M. Higdon

To cite this article: David M. Higdon (1998) Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications, Journal of the American Statistical Association, 93:442, 585-595

To link to this article: <http://dx.doi.org/10.1080/01621459.1998.10473712>



Published online: 17 Feb 2012.



Submit your article to this journal [↗](#)



Article views: 139



Citing articles: 7 View citing articles [↗](#)

# Auxiliary Variable Methods for Markov Chain Monte Carlo With Applications

David M. HIGDON

Suppose that one wishes to sample from the density  $\pi(x)$  using Markov chain Monte Carlo (MCMC). An auxiliary variable  $u$  and its conditional distribution  $\pi(u|x)$  can be defined, giving the joint distribution  $\pi(x, u) = \pi(x)\pi(u|x)$ . A MCMC scheme that samples over this joint distribution can lead to substantial gains in efficiency compared to standard approaches. The revolutionary algorithm of Swendsen and Wang is one such example. Besides reviewing the Swendsen–Wang algorithm and its generalizations, this article introduces a new auxiliary variable method called *partial decoupling*. Two applications in Bayesian image analysis are considered: a binary classification problem in which partial decoupling outperforms Swendsen–Wang and single-site Metropolis methods, and a positron emission tomography (PET) reconstruction that uses the gray level prior of Geman and McClure. A generalized Swendsen–Wang algorithm is developed for this problem, which reduces the computing time to the point where MCMC is a viable method of posterior exploration.

KEY WORDS: Please supply.

## 1. INTRODUCTION

The introduction of auxiliary variables to a Markov chain Monte Carlo (MCMC) scheme (Besag and Green 1993; Edwards and Sokal 1988) may allow one to construct Markov chains that mix faster and are easier to simulate than standard single site algorithms. The idea was given formally in the aforementioned references but was alluded to in the ideas of conditional simulation advanced by Hammersley (1956) and Trotter and Tukey (1956). In the method of auxiliary variables, one seeks to generate realizations from a complicated distribution with density  $\pi(x)$ . The variable of interest  $x \in \mathcal{X}$  is augmented by one or more additional variables  $u \in \mathcal{U}$ ; in some contexts  $u$  may have a physical interpretation in the original process, such as temperature or an unobserved measurement, though this is not necessary. To generate realizations from  $\pi(x)$ , one specifies the conditional distribution  $\pi(u|x)$  and writes  $\pi(x, u) = \pi(x)\pi(u|x)$  with marginal distribution  $\pi(x)$ . A Markov chain is then constructed on  $\mathcal{X} \times \mathcal{U}$  by alternately updating  $u$  and  $x$  via Gibbs sampling or some other method that maintains  $\pi(x, u)$ , and hence  $\pi(x)$ . Note that lower case  $x$  and  $u$  are used to represent random and standard variables.

The general auxiliary variable method may be implemented as follows:

1. Specify  $u$  and conditional distribution  $\pi(u|x)$ .
2. Form joint distribution  $\pi(x, u) = \pi(x)\pi(u|x)$ .
3. Define transition kernels  $P_u((x, u) \rightarrow (x, u'))$  and  $P_x((x, u) \rightarrow (x', u))$  such that both kernels maintain  $\pi(x, u)$ . Typically,  $u$  is updated with a Gibbs step,

$$P_u((x, u) \rightarrow (x, u')) = \pi(u'|x),$$

and  $P_x$  is specified so that detailed balance is maintained,

$$\pi(x, u)P_x((x, u) \rightarrow (x', u)) = \pi(x', u)P_x((x', u) \rightarrow (x, u)).$$

4. Generate realizations  $(x^1, u^1), \dots, (x^N, u^N)$  via the systematic scan transition kernel  $P_x P_u$  or some other updating schedule.

This article gives examples of auxiliary variable algorithms in Section 2 that show how auxiliary variable methods can lend insight into simple MCMC algorithms, and how they can lead to improvements over standard algorithms. In particular, I discuss the Swendsen–Wang algorithm (Swendsen and Wang 1987) for Ising (1925) models and introduce partial decoupling, a modification that is more relevant to statistical inference. In Section 3 two applications in Bayesian image analysis are presented where standard MCMC algorithms are impractical because they move so slowly through the sample space.

## 2. EXAMPLES

### 2.1 Metropolis

Auxiliary variables lend insight into the algorithm of Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), which can be regarded as an auxiliary variable algorithm itself. As an example, I consider a Markov chain defined on the set  $B = \{x_0, \dots, x_n\}$ , where the transition probability matrix (TPM)  $P(x_i \rightarrow \cdot)$  is a symmetric probability mass function centered at  $x_i$  with support on  $\{x_{i-2}, \dots, x_{i+2}\}$ , as shown in Figure 1. If  $B$  is periodic, then one-step transitions between  $x_n$  or  $x_{n-1}$  and  $x_0$  are possible. In this case the stationary distribution is trivially uniform.

But when  $B$  has edges, the point  $x_0$ , for example, can no longer be reached from  $x_n$  or  $x_{n-1}$ . In addition, the TPM  $P$  allows transitions outside the set  $B$ . To maintain the uniform distribution over  $B$  obtained in the aperiodic case, the Metropolis algorithm allows the chain to occasionally remain where it is for  $x_i$  near the boundary. This compensates for the loss of transitions over the boundary that occurred in the periodic case. A schematic is given in

David Higdon is Assistant Professor, Institute of Statistics and Decision Sciences, Duke University. The research was supported by National Science Foundation grants DMS 9505114 and DMS 9704425. The author would like to thank Julian Besag for stimulating conversation on Markov chain Monte Carlo. The author is also grateful to Adrian Raftery for providing the ice floe data and to Valen Johnson for providing the PET data.

Figure 1. Here the uniform Metropolis kernel is defined by

$$P_{MU}(x_i \rightarrow x_j) = \begin{cases} P(x_i \rightarrow x_i) + \sum_{x_k \notin B} P(x_i \rightarrow x_k) & \text{for } i = j \\ P(x_i \rightarrow x_j) & \text{for } i \neq j. \end{cases} \quad (1)$$

The term  $\sum_{x_k \notin B} P(x_i \rightarrow x_k)$  causes the chain to remain at  $x_i$  whenever  $P$  proposes a value  $x_k \notin B$ . To see that this mimics transitions in the periodic case, consider  $x_i$  near the boundary and  $x_k \notin B$ . In the case where  $B$  has edges,  $P(x_i \rightarrow x_k) = P(x_k \rightarrow x_i)$ , which is equal to  $P(x_{k \bmod n+1} \rightarrow x_i)$  when  $B$  is periodic. Hence each term  $P(x_i \rightarrow x_k)$  with  $x_k \notin B$  mimics the term  $P(x_{k \bmod n+1} \rightarrow x_i)$  from the periodic sampler. This simple example readily extends to continuous  $B$  with any symmetric transition kernel  $P$ .

To demonstrate how Metropolis works for arbitrary densities  $\pi(x)$ , first consider an auxiliary variable Gibbs sampler for  $\pi(x)$ . Defining  $u|x \sim U[0, \pi(x)]$  leads to a uniform joint distribution

$$\pi(x, u) \propto I[(x, u) : 0 \leq u \leq \pi(x)],$$

whose marginal density for  $x$  is  $\pi(x)$ . A Gibbs sampler for this distribution consists of two uniform updates:

1.  $u|x \sim U[0, \pi(x)]$ .
2.  $x|u \sim U\{x: \pi(x) \geq u\}$ .

This auxiliary variable scheme, often called slice sampling, has been shown by Damien, Wakefield, and Walker (1997) to be quite useful in a number of applications. If  $\pi(x)$  is not readily invertible, then an alternative is to use an adaptive approach to sample from  $\pi(x|u)$ .

The Metropolis algorithm simply replaces the foregoing step 2 with an edge-correction update for uniform distributions  $P_{MU}$ . Hence the resulting updates are as follows:

1.  $u|x \sim U[0, \pi(x)]$ .
2. Replace  $x$  by  $x'$  sampled from  $P_{MU}(x \rightarrow x')$  with  $B_u = \{x: \pi(x) \geq u\}$ :
  - a. Sample  $x^c$  from the symmetric kernel  $P(x \rightarrow x^c)$ .
  - b. Set

$$x' = \begin{cases} x^c & \text{if } x^c \in \{x: \pi(x) \geq u\} \\ x & \text{if } x^c \notin \{x: \pi(x) \geq u\}. \end{cases}$$

A candidate  $x^c$  is drawn from the symmetric kernel  $P(x \rightarrow \cdot)$ ; the chain moves to  $x^c$  if  $x^c$  is in the set  $B_u = \{x: \pi(x) \geq u\}$  and remains at  $x$  if  $x^c$  falls outside of  $B_u$ . On reordering the steps, one recognizes this as the standard Metropolis algorithm for  $\pi(x)$ , as described in the statistical literature (see, e.g., Besag, Green, Higdon, and Mengersen 1995 or Tierney 1994).

## 2.2 The Swendsen-Wang Algorithm

The advent of the Swendsen-Wang (SW) algorithm for Ising and Potts models has led to a number of more general auxiliary variable methods to combat slow mixing in lattice

models from statistical physics. The original SW algorithm was designed to speed up simulation of very large Ising models near criticality. This algorithm uses auxiliary bond variables and provides a simple means of moving through the state space in a way that cannot be done with single-site updating. It is also applicable when the model contains a likelihood term as well as when the neighbor interactions are edge-dependent.

Because of the SW algorithm's success with Ising and Potts models, much effort has been spent on generalizing it to a wider class of models. Examples include continuous-spin models (De Meo and Oh 1992; Wolff 1989) and gray-level imaging (Green 1992). Here I describe the general form of the SW algorithm due to Edwards and Sokal (1988).

Suppose that the distribution of interest  $\pi(x)$  can be written in the form

$$\pi(x) \propto \pi_0(x) \prod_k b_k(x), \quad (2)$$

where  $\pi_0(x)$  is a simple distribution, perhaps with independence for the components of  $x$ . By specifying  $u = (u_1, \dots, u_k)$  and its conditional distribution,

$$\pi(u|x) = \prod_k \frac{1}{b_k(x)} I[0 \leq u_k \leq b_k(x)], \quad (3)$$

one can knock out interactions among the components of  $x$ . Given  $x$ , the components of  $u$  are independent, with each  $u_k \sim U[0, b_k(x)]$ . The distribution of  $x$  given  $u$  is then

$$\begin{aligned} \pi(x|u) &= \pi_0(x) \prod_k b_k(x) \frac{1}{b_k(x)} I[0 \leq u_k \leq b_k(x)] \\ &= \pi_0(x) \prod_k I[u_k \leq b_k(x)]. \end{aligned} \quad (4)$$

Now  $x|u$  is distributed according to  $\pi_0(x)$ , subject to the constraints,  $u_k \leq b_k(x)$  for all  $k$ . In Bayesian image analysis,  $\pi_0(x)$  typically denotes the likelihood component of the posterior distribution. For simple Ising and Potts models, the constraints are quite simple to deal with, as shown in the following example. But in general, the constraints can make simulating from  $\pi(x|u)$  a difficult task.

**2.2.1 Swendsen-Wang for Binary Markov Random Fields.** Let  $S$  denote the set of pixels or lattice sites, with pairwise adjacencies indicated by  $i \sim j$ . On a finite lattice consisting of  $n$  sites, the distribution of the Ising model with a likelihood or external field term can be written as

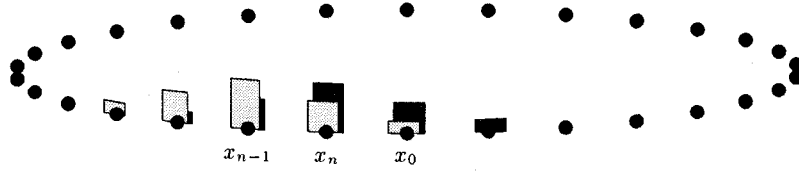
$$\pi(x) \propto \exp \left\{ \sum_{i \in S} \alpha_i(x_i) \right\} \times \exp \left\{ \sum_{i \sim j} \beta_{ij} I[x_i = x_j] \right\}, \quad (5)$$

$x \in \{0, 1\}^n$ .

Note that the Potts (1952) model has identical form, except that each  $x_i$  may take on more than two unordered states. I assume that  $\beta_{ij} > 0$ , so this distribution invites clustering of like-colored pixels. In the case of no external field ( $\alpha_i(\cdot) \equiv 0$  for all  $i$ ), I refer to (5) as the Ising model.

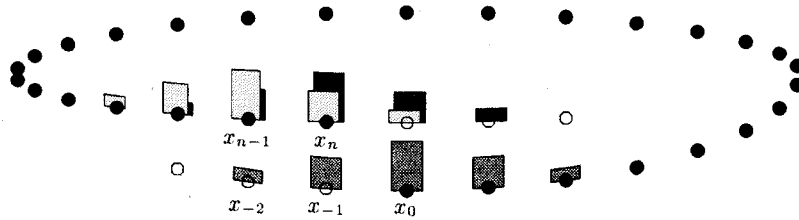
In applying SW, we define  $u$  with components  $u_{ij}$  corresponding to each adjacency  $i \sim j$  in the lattice. Given

With periodic  $B$ , one-step transitions from  $x_n$  or  $x_{n-1}$  to  $x_0$  are possible.



Here the stationary distribution using transition kernel  $P$  is trivially uniform.

When  $B$  is no longer periodic, points near the edge lose contribution from points beyond the edge.



At  $x_0$ , the Metropolis kernel  $P_{MU}$  mimics the contribution from  $x_n$  and  $x_{n-1}$  in the periodic sampler by sticking at  $x_0$  when  $P$  proposes transitions to  $x_{-1}$  or  $x_{-2}$ :

$$P_{MU}(x_0 \rightarrow x_j) = \begin{cases} P(x_0 \rightarrow x_0) + \sum_{x_k \notin B} P(x_0 \rightarrow x_k) & \text{for } x_j = x_0 \\ P(x_0 \rightarrow x_j) & \text{for } x_j \neq x_0 \end{cases}$$

Figure 1. Metropolis on a Uniform Distribution Acts as a Method for Correcting Edge Effects. Consider the space  $B = \{x_0, \dots, x_n\}$  given by the black dots in the two diagrams above and TPM  $P(x \rightarrow \cdot)$  which is a symmetric PMF centered at  $x$ . If the space  $B$  is periodic (a), then  $x_0$  receives support from  $x_n$  and  $x_{n-1}$ . However when  $B$  has edges (b), a transition from  $x_n$  or  $x_{n-1}$  to  $x_0$  is no longer possible. To maintain the uniform distribution over  $B$ , the Metropolis kernel mimics support from  $x_n$  or  $x_{n-1}$  in the periodic case by sticking at  $x_0$  when a realization from  $P(x_0 \rightarrow \cdot)$  hits  $x_{-1}$  or  $x_{-2}$ .

$x$ , the components  $u_{ij}$  are then specified to be independent and uniformly distributed,

$$\pi(u_{ij}|x) = \exp\{-\beta_{ij}I[x_i = x_j]\} \times I[0 \leq u_{ij} \leq \exp\{\beta_{ij}I[x_i = x_j]\}], \quad (6)$$

so that

$$\pi(x|u) \propto \exp\left\{\sum_{i \in S} \alpha_i(x_i)\right\} \times \prod_{i \sim j} I[0 \leq u_{ij} \leq \exp\{\beta_{ij}I[x_i = x_j]\}]. \quad (7)$$

If  $u_{ij} > 1$ , then  $\exp\{\beta_{ij}I[x_i = x_j]\} > 1$ , so the condition  $u_{ij} > 1$  implies that  $x_i = x_j$ . Like-colored neighbors,  $i$  and  $j$ , are bonded (i.e., constrained to be equal) with probability  $1 - \exp\{-\beta_{ij}\}$ . The bond variable  $u$  partitions  $S$  into like-colored clusters. For a particular cluster,  $C$ , the probability of color  $k \in \{0, 1\}$  is  $\propto \exp\{\sum_{i \in C} \alpha_i(k)\}$ , so each cluster can be updated independently according to its conditional distribution. The bond variables completely decouple the external field term,  $\exp\{\sum_{i \in S} \alpha_i(x_i)\}$ , from the interaction term,  $\exp\{\sum_{i \sim j} \beta_{ij}I[x_i = x_j]\}$ . Updating  $u$  essentially grows clusters, and updating  $x$  colors them. Figure 2 illustrates the SW algorithm on an  $8 \times 8$  lattice with first-

order neighborhood structure and no external field. Figure 3 shows successive realizations of the SW algorithm and Gibbs sampling on a  $100 \times 100$  lattice with all  $\beta_{ij}$  set to the critical value  $\beta^*$  for the infinite lattice,  $\beta^* \approx .88$ .

**2.2.2 Swendsen-Wang for a Gray-Level Model.** For the Ising model, SW gives the most improvement of single-site Metropolis when  $\beta$  is at the critical value, so that realizations are patchy as in Figure 3. This suggests that SW may give substantial improvement for gray-level priors that also yield patchy realizations. Although the commonly used Gaussian pairwise difference prior (Besag et al. 1995, sec. 3) does not exhibit this behavior, the prior model of Geman and McClure (1987),

$$\pi(x|\beta, \lambda) \propto \exp\left\{\beta \sum_{i \sim j} [1 + \lambda(x_i - x_j)^2]^{-1}\right\}, \quad x \in [0, M]^n, \quad (8)$$

does show some similarities to the Ising model, which suggests that SW may be effective in sampling from such a model. Figure 4 shows every 20th realization from (8) after reaching the stationary distribution under a single-site Metropolis algorithm and a SW implementation with  $(\beta, \lambda, M) = (.96, .005, 300)$ . These parameter values result

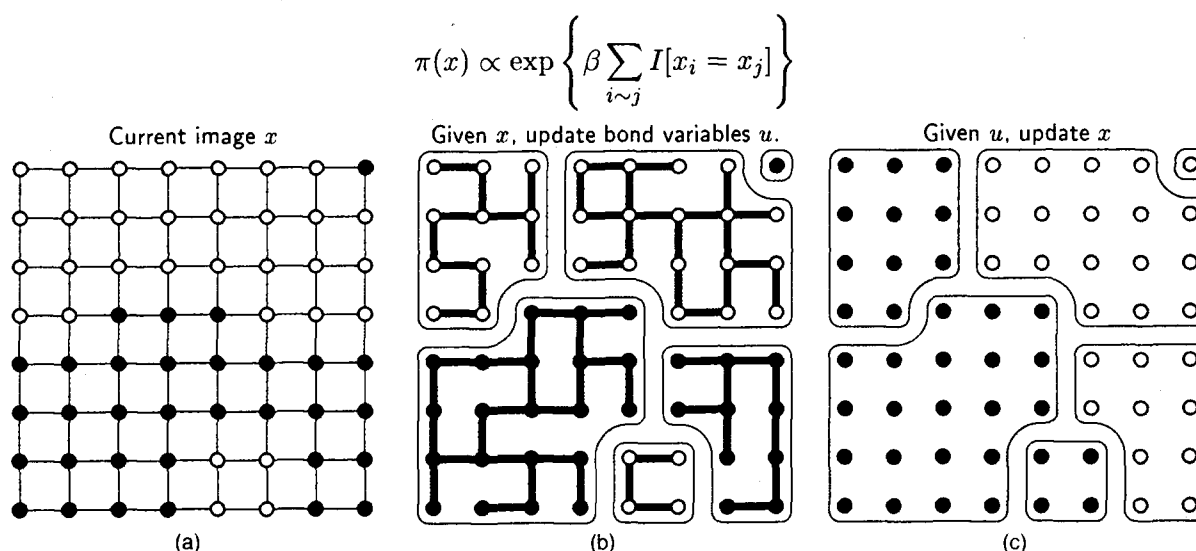


Figure 2. The Swendsen-Wang Algorithm for the Ising Model on the  $8 \times 8$  Lattice. (a) Initial image  $x$  and Markov random field graph for  $\pi(x)$ . (b) Given the current image  $x$ , the bond variables  $u$  are generated uniformly over the interval  $(0, e^{\beta I[x_i = x_j]})$ . If  $u_{ij} > 1$  (marked by the thick lines),  $x_i$  is constrained to equal  $x_j$ . These constraints partition the image into clusters of like-colored sites. Clusters induced by this realization of  $u|x$  are outlined. The Markov random field graph for  $x|u$  differs from that of  $x$ , marginally; the auxiliary vector  $u$  strengthens the dependence between some neighboring sites, while completely removing it from others. (c) Given the bond variables  $u$ ,  $x$  is now a coarse image of independent clusters. Because there is no external field in this example, each cluster is recolored black or white with probability .5.

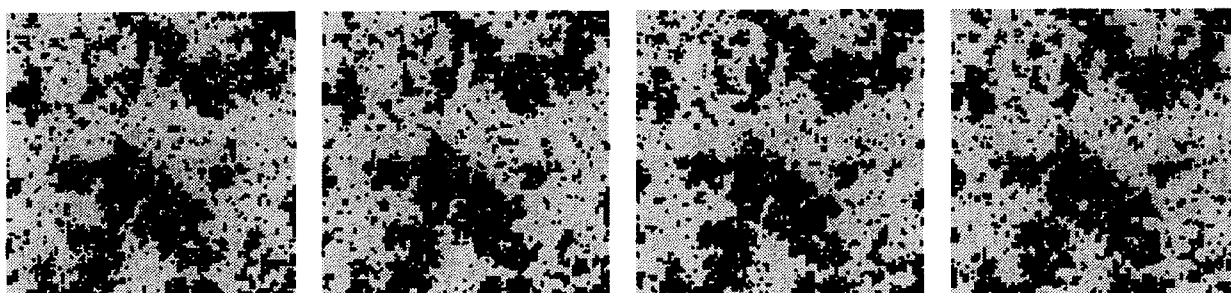
in patchy realizations for  $x$ . As in standard SW applied to the Ising model, conditioning on  $u$  leads to independence between clusters and dependence within clusters through the cumbersome constraints (4). Rather than attempt to update within a cluster, we consider shifting the level of the cluster, leaving the relative pairwise differences within each cluster unchanged. The level of the cluster follows a uniform distribution, though some care must be taken that all

of the  $x_i$ 's within a cluster remain in the interval  $[0, M]$  when updating. Thus the SW implementation for (8) can be described as follows:

1. Update each bond variable according to a uniform distribution:

$$u_{ij}|x \sim U[0, \exp\{\beta[1 + \lambda(x_i - x_j)^2]^{-1}\}].$$

### successive realizations



Gibbs

Swendsen - Wang

Figure 3. Successive Realizations of the Ising Model at Critical Temperature From the Single-Site Metropolis and Swendsen-Wang Algorithms.

every 20th realization

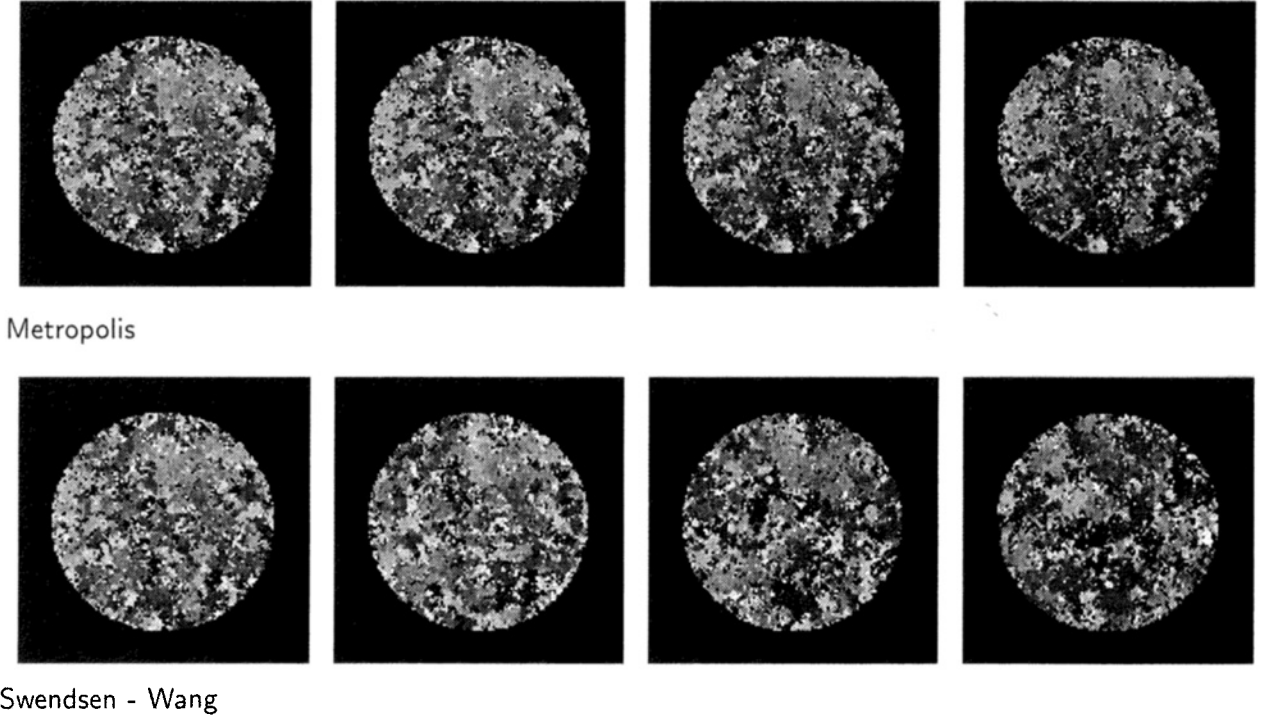


Figure 4. Successive Realizations of the Geman and McClure Prior at "Critical Temperature" From the Single-Site Metropolis and the Generalized Swendsen-Wang Algorithms.

2. Determine clusters  $x_{C_1}, \dots, x_{C_n}$  induced by bond variable  $u$ .

3. Replace each  $x_{C_i}$  by  $x_{C_i} + r_i$ , where  $r_i$  is a single uniform draw from the interval  $[-\min(x_{C_i}), M - \max(x_{C_i})]$ .

Although this algorithm results in an irreducible Markov chain, I alternate the SW step above with a single-site Metropolis step to ensure movement within large patches. Figure 5 shows a time series plot of  $u(x) = \sum_{i \sim j} [1 + \lambda(x_i - x_j)^2]^{-1}$  and the mean level of  $x$  by iteration. Because the SW algorithm alternates with single-site Metropolis sweeps, values at every second iteration are plotted for the single-site Metropolis algorithm to make a fairer comparison. This new SW implementation is a huge improvement over single-site Metropolis and is key in the example of Section 3.2, where required draws from both the prior and posterior distributions would be impossible to obtain with single-site methods.

### 2.3 The Partial Decoupling Method

The method of *partial decoupling* was first given by Higdon (1993). The algorithm can be laid out in a slightly more general form by replacing (3) with

$$\pi(u|x) = \prod_k \frac{1}{b_k(x)^{\delta_k}} I[0 \leq u_k \leq b_k(x)^{\delta_k}],$$

leading to the conditional distribution of  $x|u$ ,

$$\pi(x|u) = \pi_0(x) \prod_k b_k(x)^{1-\delta_k} I[u_k \leq b_k(x)^{1-\delta_k}].$$

This method has been used successfully for sampling from posterior distributions resulting from binary imaging applications. In such problems the likelihood term breaks the symmetry of the prior model, so that it is typically inefficient to allow clusters to grow without regard to the likelihood and then to allow the clusters to be updated without regard to the prior as occurs in SW. In partial decoupling,  $\pi(u|x)$  is modified so that the prior is only partially decoupled from the likelihood term when considering  $\pi(x|u)$ . In what follows I give the details of applying partial decoupling to binary Markov random fields.

**2.3.1 Partial Decoupling for Binary Markov Random Fields.** The method of partial decoupling was originally developed for binary image applications that result in a posterior distribution of the form (5). Set

$$\pi(u_{ij}|x) = \exp\{-\delta_{ij}\beta_{ij}I[x_i = x_j]\} \times I[0 \leq u_{ij} \leq \exp\{\delta_{ij}\beta_{ij}I[x_i = x_j]\}],$$

which yields

$$\pi(x|u) \propto \exp\left\{\sum_{i \in S} \alpha_i(x_i) + \sum_{i \sim j} (1 - \delta_{ij})\beta_{ij}I[x_i = x_j]\right\} \times \prod_{i \sim j} I[0 \leq u_{ij} \leq \exp\{\delta_{ij}\beta_{ij}I[x_i = x_j]\}].$$

The term  $\prod_{i \sim j} I[0 \leq u_{ij} \leq \exp\{\delta_{ij}\beta_{ij}I[x_i = x_j]\}]$  induces clusters as in the SW algorithm, but here like-colored neighbors  $i$  and  $j$  bond with probability  $1 - \exp\{-\delta_{ij}\beta_{ij}\}$ . Conditional on the bond variables  $u$ , the clusters are no longer

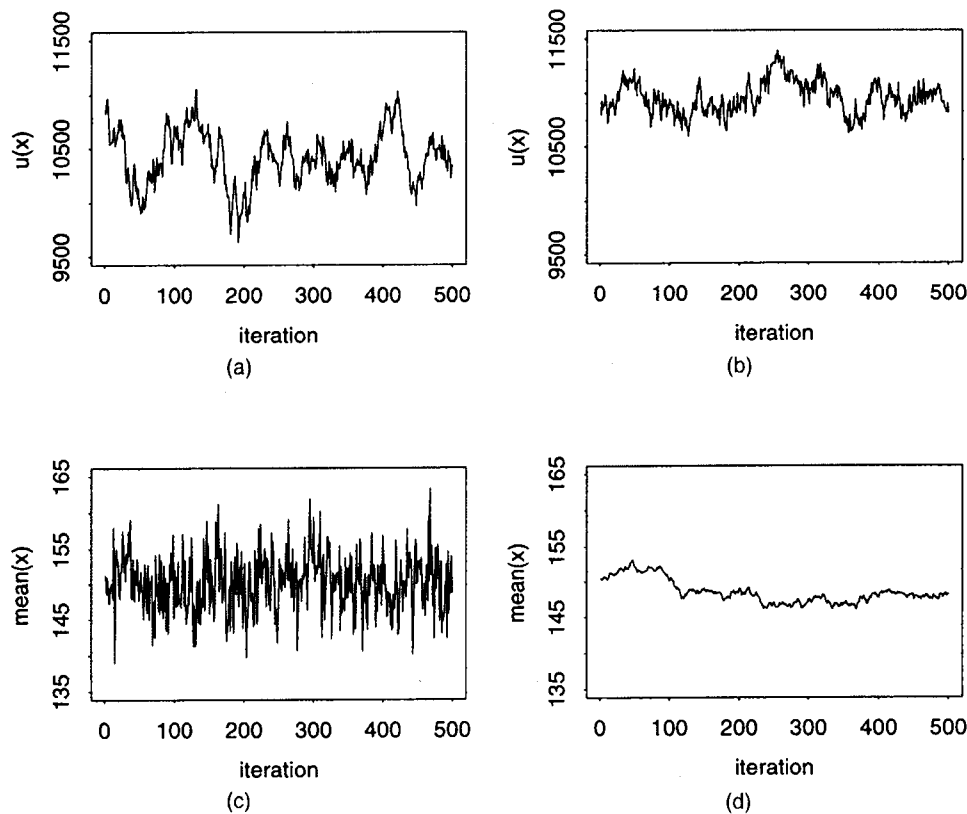


Figure 5. Comparison of Time Series Plots of  $u(x) = \sum_{i \sim j} [1 + \lambda(x_i - x_j)^2]^{-1}$  and  $\sum x/n$  From MCMC Simulation of the Geman-McClure Prior Using Swendsen-Wang [(a), (c)] and Single-Site Metropolis [(b), (d)] Updating.

independent of one another, and hence it is necessary to update clusters conditionally on their neighbors. Note that  $\pi(x|u)$  is a stochastically coarsened version of the original Ising model; updating  $x|u$  can be done sitewise by cluster via a Gibbs, a Metropolis, or an auxiliary variable method for this coarsened model. The bonding probability is controlled by the  $\delta_{ij}$ 's; note that all  $\delta_{ij} = 0$  corresponds to Gibbs sampling, whereas all  $\delta_{ij} = 1$  corresponds to the SW algorithm. Figure 6 illustrates the partial decoupling algorithm. For simplicity, all  $\delta_{ij} = \delta$ ,  $0 < \delta < 1$ , and no external field is considered. Again, the image resides on an  $8 \times 8$  lattice with first-order neighborhood structure.

**2.3.2 Choosing  $\delta$ .** In practice, one is left to specify the constants  $\delta_{ij}$ . The basic strategy behind any choice is to grow clusters that improve mixing when updating  $x|u$ . Depending on the nature of the likelihood for  $x$ , numerous strategies may be used. If certain components of  $x$  are fixed, or nearly fixed by the likelihood, then SW updating can fare very poorly. The problem arises because most of the sites belong to a cluster that contains one or more of the fixed sites. Because the cluster contains the fixed site, the probability of the cluster changing color is 0. Hence most of the image remains unchanged after each iteration. If  $x_i$  is fixed at 0, for instance, then setting  $\delta_{ij} = 0$  for all  $j \in \partial i$ , and leaving the remaining  $\delta$ 's at 1 will keep any neighboring site from bonding with the fixed site, while using the SW scheme away from the fixed site (Higdon 1994 for an example).

Another setting in which the SW algorithm can lead to slow mixing is when the external field term of the Ising model results in a multimodal distribution for  $x$ . The original SW algorithm fares very well when there is no external field to break stationarity and symmetry properties of the distribution. Because the clusters form without regard to the external field term in the SW algorithm, the clusters that form may have very little chance of changing in the presence of the external field. Perhaps the simplest approach is to set  $\delta_{ij} = \delta$ . Because  $\delta$  controls the chance that neighboring pixels will bond, setting  $\delta$  to a sufficiently small constant will ensure that clusters do not grow too large. An alternative is to choose the  $\delta_{ij}$ 's to prevent clusters from growing too large or across certain boundaries. For example, the lattice may be broken into sublattices. Setting  $\delta_{ij} = 0$  for adjacencies linking sublattices, while leaving the remaining  $\delta$ 's at 1, prevents clusters from growing beyond these boundaries. Conditional on the bond variable  $u$ , these clusters will no longer be independent as they were in the standard SW. Note that one may change the values for  $\delta$  at each iteration according to some deterministic schedule. (See Hurn 1997 for an application that adaptively blocks the array into successively coarser sublattices.)

Perhaps the most appealing method for determining  $\delta$  in imaging applications is by considering the likelihood. Often the likelihood is a function of the absolute difference between the data records  $y$  and the restored image  $x$ . For adjacent pixels  $i$  and  $j$ , choose  $\delta_{ij}$  to be near 1 if  $y_i$  and  $y_j$  are similar; choose  $\delta_{ij}$  to be near 0 if  $y_i$  and  $y_j$  are disparate.

This strategy gives clusters that are more likely to change when updating  $x|u$ . The choice of  $\delta_{ij} = \phi(|y_i - y_j|)$ , where  $\phi(\cdot)$  is a decreasing function ranging from 1 to 0, has proven fruitful in a number of binary imaging applications. With the binary records in Section 3.1, I use  $\phi(u) = aI[u = 0]$ , where  $a$  is a chosen constant. In cases where the data records are gray levels,  $\phi(u) = (1 + |u|)^{-1}$  works well in the example of Higdon (1993).

### 3. APPLICATIONS

This section considers two applications from Bayesian image analysis using the Ising prior and the gray level prior of Geman and McClure (1986) discussed earlier. The application first is binary classification problem in which a simple Ising prior is used to aid in the identification of ice floes. The resulting posterior distribution is multimodal, and posterior exploration via standard Metropolis or SW gives misleading results. Partial decoupling seems to avoid the pitfalls of the other MCMC schemes. The second application is a positron emission tomography (PET) reconstruction of cerebral blood flow. The fully Bayesian analysis conducted would be infeasible without making heavy use of the SW algorithm constructed in Section 2.2 for the Geman–McClure prior. The sampling scheme allows the interaction parameter to vary throughout the MCMC run. Both unknown normalizing constants and critical behavior of these two priors pose difficulty in implementation and require simulation via the cluster algorithm described in Section 2.2.

#### 3.1 Identification of Ice Floes

To highlight partial decoupling, an application from Banfield and Raftery (1992) is considered. Here the goal is to identify ice floes in a polar LANDSAT image. As suggested

in their discussion, I take a Bayesian imaging approach, basing the likelihood on the original image  $y$  and assigning a simple Ising model prior to the unknown binary image  $x$ . As in the original application, the LANDSAT image is thresholded so that the data records  $y$  are also binary, as shown in Figure 7a. distribution is

$$\pi(x|y) \propto \exp \left\{ \alpha \sum_i I[y_i = x_i] + \beta \sum_{i \sim j} I[x_i = x_j] \right\}$$

for  $x \in \{0, 1\}^{194 \times 200}$ . The spatial prior for  $x$  uses horizontal, vertical, and diagonal adjacencies, so that each interior pixel has eight neighbors.

The parameters  $(\alpha, \beta)$  are fixed at  $(1, .8)$ . The value for the Ising parameter  $\beta$  is specified so that the posterior realizations  $x$  will contain very little speckle, whereas the value for  $\alpha$  is set so that the prior influence will polish large patches of ice and remove small, irregular patches that are present in the data records  $y$ . Although these parameter values lead to appealing realizations from  $\pi(x|y)$ , the posterior is multimodal—occasional smaller patches are present in some realizations and absent in others. This is exactly the situation in which single-site Metropolis and SW perform very poorly. Figure 7b shows the estimated posterior mean for  $x$  from a partial decoupling run of 40,000 iterations.

I tried a number of partial decoupling schemes on this posterior distribution. Of these, the most successful uses a simple recipe for which  $\delta_{ij} = aI[y_i = y_j]$ . At  $a = 0$ , the algorithm is single-site Metropolis; at  $a = 1$ , it is a “blocked” SW scheme, where bonds may not form between adjacent sites  $i$  and  $j$  if  $y_i \neq y_j$ . To evaluate performance at various  $a$ , I focus on a subset of the image marked by the  $20 \times 20$  square outlined in Figure 7b. This square  $S$  is one of the

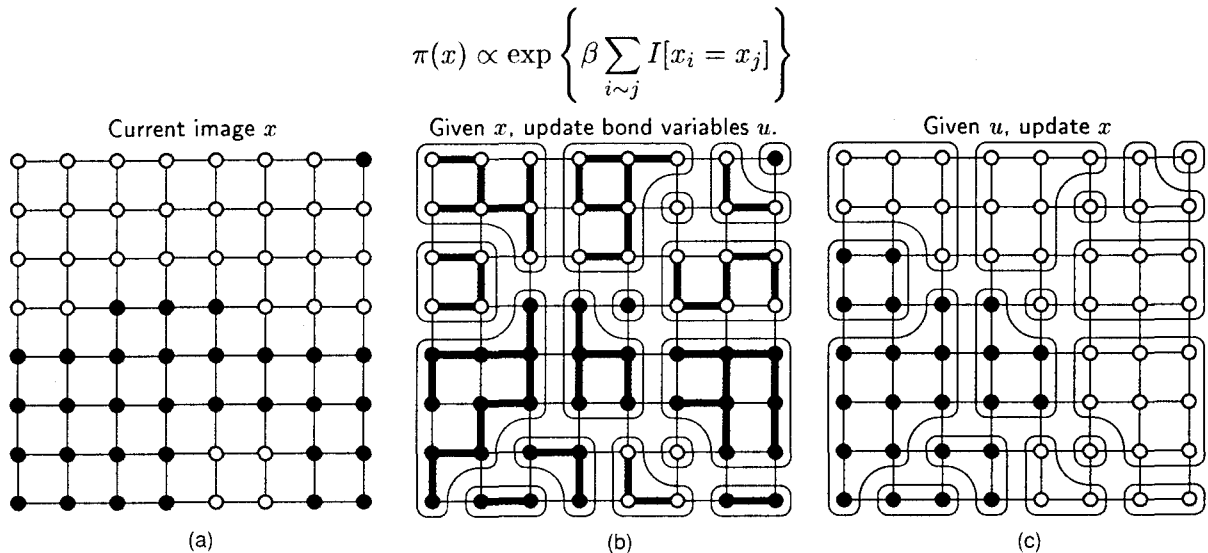


Figure 6. The Partial Decoupling Algorithm With  $0 < \delta_{ij} = \delta < 1$  for the Ising Model on the  $8 \times 8$  Lattice. (a) Initial image  $x$  and Markov random field graph for  $\pi(x)$ . (b) Given the current image  $x$ , the bond variables  $u$  are generated uniformly over the interval  $(0, e^{\delta \beta I[x_i = x_j]})$ . If  $u_{ij} > 1$  (marked by the thick lines),  $x_i$  is constrained to equal  $x_j$ . In the partial-decoupling scheme,  $\Pr(u_{ij} > 1)$  is smaller than in Swendsen–Wang, so these constraints are less likely to form. As in Swendsen–Wang, the constraints partition the image into clusters of like-colored sites. Clusters induced by this realization of  $u|x$  are outlined. The Markov random field graph for  $x|u$  is also different; although the auxiliary vector  $u$  may still lead to constraints, the dependence between neighboring sites is never completely removed. (c) Because the bond variable  $u$  does not remove dependence between clusters,  $\pi(x|u)$  is now a coarsened version of the original Ising model. Each cluster is updated conditionally on its neighboring clusters. This coarser Ising model may be updated via Gibbs, Metropolis, or even partial-decoupling again.





## Positron Emission Tomography (PET)

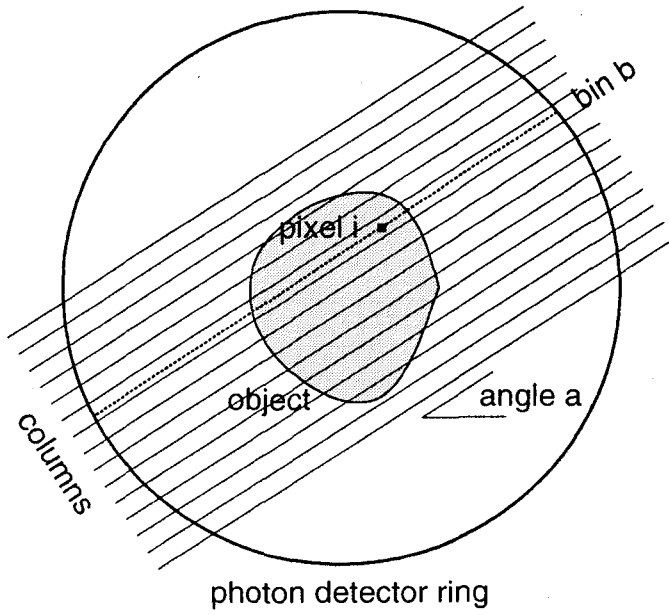


Figure 8. PET Reconstructions. A pixilated object emits positrons with location-dependent intensities  $x_i$ . Immediately after a positron is emitted from pixel  $i$ , an annihilation event causes two photons to be emitted in directly opposite directions, so that pixel  $i$  must be contained within a thin column determined by the photon paths. Each pair of "hits" corresponds to a column indexed by angle of the column  $a$  and the bin number  $b$ , giving binned counts  $y_{ab}$ . Because a photon may be scattered, absorbed, miss the gamma camera, or otherwise fail to be detected, the probability map  $p_{abi}$  gives the probability of an emission from pixel  $i$  being detected at angle  $a$  and column bin  $b$ .

Because  $Z(\beta)$  is analytically intractable, prior simulation via MCMC is required to estimate  $Z(\beta)$  over  $S$  using reverse logistic regression (Geyer 1991, 1997). Alternatively, thermodynamic integration (Ogata and Tanemura 1984) or path sampling (Gelman and Meng 1996) could have been used. All of these methods require draws from the prior (8), which could not be done in a reasonable amount of time without the SW algorithm devised in Section 2.2.

The resulting posterior distribution is

$$\pi(x, \beta | y) \propto \prod_{ab} \mu_{ab}^{y_{ab}} e^{-\mu_{ab}} \frac{1}{Z(\beta)} \times \exp \left\{ \beta \sum_{i \sim j} [1 + \lambda(x_i - x_j)^2]^{-1} \right\},$$

$$x \in [0, M]^n, \quad \beta \in S.$$

The full conditional for  $x$  is handled using the generalized SW algorithm. Because of the dependence between components of  $x$  induced by the PET likelihood term, sampling directly from  $\pi(x|u)$  is impossible. Instead, each cluster is updated conditional on the current value of all  $x_i$ 's. Once the clusters have been determined, a separate Metropolis proposal is made to adjust the overall level of the cluster. For a given cluster  $x_C$ , a symmetric proposal is made,  $x'_C = x_C + v$ , where  $v$  is a  $U[-110, 110]$  random variable divided by the square root of the size of the cluster. Provided that all components of  $x'_C$  are between 0 and  $M$ , the proposal is then accepted with probability

$$\min\{1, L(x'|y)/L(x|y)\},$$

where  $x'$  is identical to  $x$  except that  $x'_C$  replaces  $x_C$ . This gives about a 60% acceptance rate on average. The interaction parameter  $\beta$  is updated via a Metropolis step that proposes one of the two adjacent values in  $S$  as the candidate.

Figure 9 shows selected realizations from the posterior distribution along with the posterior mean of the image  $x$ . The approximate 95% credible interval for the interaction parameter is  $[\cdot96, \cdot97]$ . Of course, this is only an initial attempt to treat  $\beta$  as a parameter in the MCMC simulation. A more thorough analysis should also consider the choices of  $\lambda$  and  $M$ , possibly treating them as dynamic parameters in the simulation.

## 4. DISCUSSION

The applications presented here show that auxiliary vari-

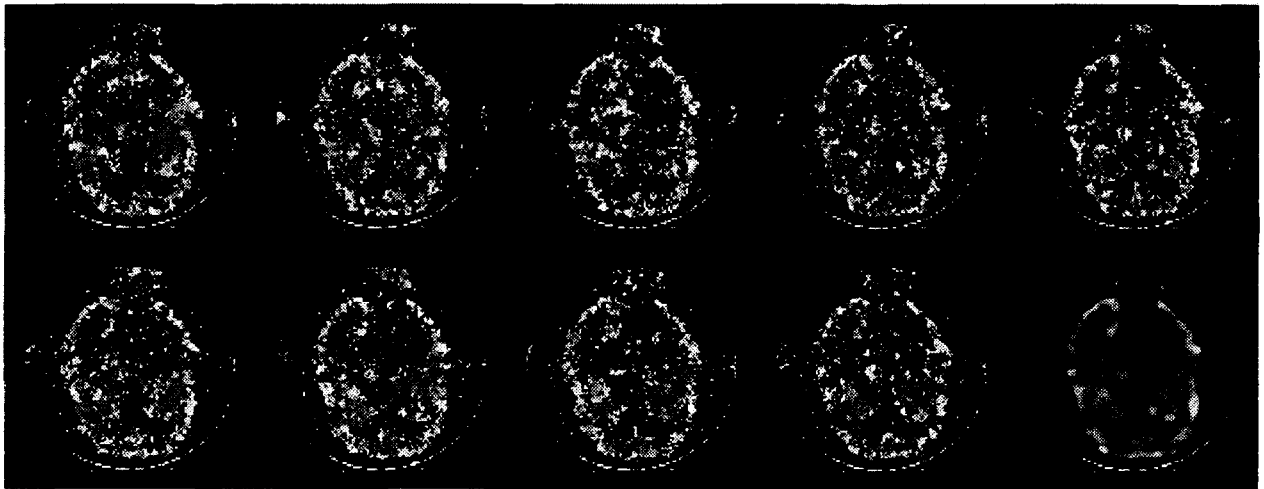


Figure 9. Realizations From the Posterior Generated Using SW Updates for the Pixel Values  $x$  and Metropolis for  $\beta$ . The final image is an estimate of the posterior mean for  $x$ .

able methods can lead to substantial gains in efficiency when MCMC is used to explore the posterior distribution resulting from an imaging problem. Success of auxiliary variables in certain Bayesian image applications is due largely to the simple dependence structure in the lattice priors. That structure is exploited so that the constraints induced by  $\pi(x|u)$  are satisfied when updating  $x|u$ . Another key in the success of auxiliary variable methods is recognizing distributions for which such algorithms are likely to fare better than standard single-site approaches. Although no clear rules as yet exist for identifying such distributions, those that show patchiness and multimodality seem to be good candidates. Both posteriors resulting from the ice floe and the PET applications show this tendency.

The Ising and GM priors (Sec. 2.2) have similarities worth noting. They both respond well to SW updating but not to single-site updating; they both are symmetric distributions without influence from a likelihood or external field; they both show a form of criticality where realizations tend to look patchy. The patchiness of the realizations has a large influence on the size of the clusters formed when updating  $u|x$ . The lack of a strong likelihood term allows rather substantial movement when updating  $x|u$ . Under such conditions, general SW algorithms likely will be quite successful.

The two applications presented in Section 3 differ in the relative strengths of their priors and likelihoods. In the ice floes example, both the Ising prior and the likelihood are quite strong. Updating via SW is ineffective, because clusters grow without regard to the likelihood term. Hence it is very unlikely that a cluster will form that can change color. Partial decoupling is effective here, because it governs the cluster size and forms clusters that facilitate movement between local modes in  $\pi(x|y)$ . In the PET example, the likelihood term has much less influence on the very local properties of  $x$ . Unlike the ice floes example, the likelihood is not independent; the effect of any  $x_i$  is "blurred" over more than 15,000 bins. Here SW fares quite well, because the full conditional distribution for clusters  $x_C$  is sufficiently spread out to allow satisfactory mixing.

Outside of lattice models, the slice sampler of Section 2.1 has been proposed as an alternative to Hastings-type algorithms for posterior distributions resulting from Bayesian models with nonconjugate formulations given by Damien et al. (1997). Although one of the main motivations for its use is ease in coding, recent work has shown that theoretical properties of slice samplers are very good (Fishman 1996; Mira and Tierney 1997; Roberts and Rosenthal 1997). In fact, Mira and Tierney (1997) showed that the slice sampler is superior to an independence Metropolis sampler. This is not surprising in light of the reexpression of the Metropolis algorithm as an auxiliary variable method in Section 2.1. Although some may be unhappy with the increase in dimensionality for slice samplers, this increase also occurs implicitly for Metropolis samplers. In addition, the slice sampler uses a Gibbs update for  $x|u$ , whereas Metropolis uses the clunky edge-correction kernel.

I conclude by pointing out that the slice sampler implementations of Damien et al. make no attempt to update  $x|u$  simultaneously when  $x$  is multidimensional. Rather, each component of  $x$  is updated in turn, conditional on all of its other components. In the two main applications considered in this article, being able to update  $x|u$  simultaneously plays a key role in developing MCMC algorithms for distributions for which single-site methods are hopelessly slow mixing.

[Received July 1996. Revised October 1997.]

## REFERENCES

- Banfield, J. D., and Raftery, A. E. (1992), "Ice Floe Identification in Satellite Images Using Mathematical Morphology and Clustering About Principal Curves," *Journal of the American Statistical Association*, 87, 7-16.
- Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 16, 395-407.
- Besag, J., Green, P. J., Higdon, D. M., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3-66.
- Damien, P., Walker, S., and Wakefield, J. (1997), "Gibbs Sampling for Bayesian Nonconjugate Models Using Auxiliary Variables," Technical Report 9705-13, University of Michigan Business School.
- Edwards, R. G., and Sokal, A. D. (1988), "Generalization of the Fortuin-Kasteleyn-Swendsen-Wang Representation and Monte Carlo Algorithm," *Physical Review Letters*, 38, 2009-2012.
- Fishman, G. S. (1996), "An Analysis of Swendsen-Wang and Related Sampling Methods," Technical Report 96-04, University of North Carolina, Dept. of Operations Research.
- Gelman, A., and Meng, X. (1996), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," Technical Report 440, University of Chicago, Dept. of Statistics.
- Geman, S., and McClure, D. (1987), "Statistical Methods for Tomographic Image Reconstruction," *Bulletin of the International Statistical Institute*, 52, 5-21.
- Geyer, C. J. (1991), "Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo," Technical Report 568, University of Minnesota, School of Statistics.
- (1997), "Likelihood Inference for Spatial Point Processes," in *Current Trends in Stochastic Geometry and Its Applications*, eds. W. S. Kendall, O. E. Barndorff-Nielsen, and M. C. van Lieshout, London: Chapman and Hall.
- Green, P. J. (1992), "A Note on the Swendsen-Wang Algorithm for Ordered Colours," Technical Report, University of Bristol, Statistics Group.
- Hammersley, J. M. (1956), "Conditional Monte Carlo," *Computing Machines*, 3, 73-76.
- Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.
- Higdon, D. M. (1993), "Comment on 'Spatial Statistics and Bayesian Computation' by J. Besag and P. Green," *Journal of the Royal Statistical Society, Ser. B*, 55, 78.
- (1994), "Spatial Applications of Markov Chain Monte Carlo for Bayesian Inference," unpublished Ph.D. dissertation, University of Washington, Dept. of Statistics.
- Hurn, M. A. (1997), "Difficulties in the Use of Auxiliary Variables in Markov Chain Monte Carlo Methods," *Statistics and Computing*, 7, 35-44.
- Ising, E. (1925), "Beitrag zur Theorie des Ferromagnetismus," *Zeitschrift für Physik*, 31, 253-258.
- Kandel, D., Romany, E., and Brandt, A. (1989), "Simulations Without Critical Slowing Down: Ising and Three-State Potts Models," *Physical Review, Ser. B*, 40, 330-344.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087-1091.
- Mira, A., and Tierney, L. (1997), "On the Use of Auxiliary Variables in Markov Chain Monte Carlo Sampling," Technical Report, University of

- Minnesota, School of Statistics.
- Ogata, Y., and Tanemura, M. (1984), "Likelihood Analysis of Spatial Point Patterns," *Journal of the Royal Statistical Society, Ser. B*, 46, 496–518.
- Potts, R. B. (1952), "Some Generalized Order-Disorder Transformations," *Proceedings of the Cambridge Philosophic Society*, 48, 106–109.
- Roberts, G. O., and Rosenthal, J. S. (1997), "Convergence of Slice Sampler Markov Chains," Technical Report, Cambridge University, Statistical Laboratory.
- Shepp, L., and Vardi, Y. (1982), "Maximum Likelihood Reconstructions for Emission Tomography," *IEEE Transactions on Medical Imaging*, 1, 113–122.
- Sokal, A. D. (1987), "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms," *Cours de Troisième Cycle de la Physique en Suisse Romande*, Lausanne.
- Swendsen, R. H., and Wang, J. S. (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physical Review Letters*, 58, 86–88.
- Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 21, 1701–1762.
- Trotter, H. F., and Tukey, J. W. (1956), "Conditional Monte Carlo for Normal Samples," in *Symposium on Monte Carlo Methods*, ed. H. A. Meyer, New York: Wiley, pp. 64–79.
- Vardi, Y., Shepp, L., and Kaufman, L. (1985), "A Statistical Model for Positron Emission Tomography," *Journal of the American Statistical Association*, 80, 8–25.