Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications

David M. Higdon *

Submitted to JASA Theory and Methods July, 1996.
Revised September, 1997.

Abstract

Suppose 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 which samples over this joint distribution can lead to substantial gains in efficiency compared to standard approaches. The revolutionary algorithm of Swendsen and Wang (1987) is one such example. In addition to reviewing the Swendsen-Wang algorithm and its generalizations, this paper introduces a new auxiliary variable method called partial decoupling. Two applications in Bayesian image analysis are considered. The first is a binary classification problem in which partial decoupling out performs SW and single site Metropolis. The second is a PET reconstruction which uses the gray level prior of Geman and McClure (1987). A generalized Swendsen-Wang algorithm is developed for this problem, which reduces the computing time to the point that MCMC is a viable method of posterior exploration.

1 Introduction

The introduction of auxiliary variables to a MCMC scheme (Edwards and Sokal, 1988; Besag and Green, 1993) may allow one to construct Markov chains which are faster mixing and easier to simulate than standard single site algorithms. The idea is given formally in the above references, but is alluded to in the ideas of conditional simulation given in Hammersley (1956) and Trotter

*David Higdon is Assistant Professor, Institute of Statistics and Decision Sciences, Duke University. The research was supported by NSF 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.

and Tukey (1956). In the method of auxiliary variables we seek 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. In order to generate realizations from $\pi(x)$, we specify the conditional distribution $\pi(u|x)$ and write $\pi(x,u) = \pi(x)\pi(u|x)$ 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 as well as 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 $P_u((x,u) \to (x,u'))$ and $P_x((x,u) \to (x',u))$ such that both kernels maintain $\pi(x,u)$. Typically, u is updated with a Gibbs step

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

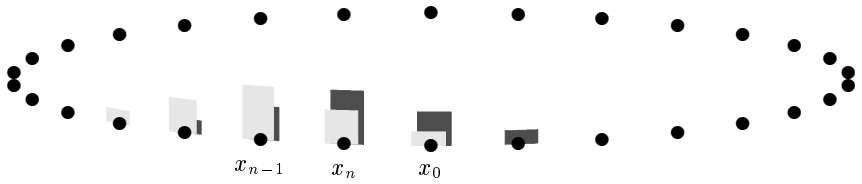
   and P_x is specified so that detailed balance is maintained

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

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

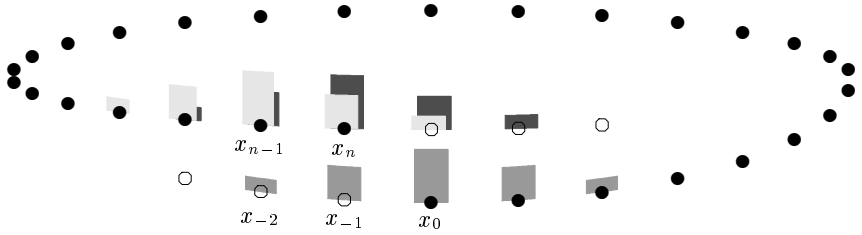
This paper gives examples of auxiliary variable algorithms in Section 2 which show how auxiliary variable methods can lend insight into simple MCMC algorithms, and how they can lead to improvements over standard algorithms. In particular, we discuss the Swendsen-Wang algorithm (Swendsen and Wang, 1987)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

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 perriodic sampler by sticking at x_0 when P proposes transitions to x_{-1} or x_{-2}:

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

Figure 1: Metropolis on a uniform distribution acts as a method for correcting edge effects. Consider the space $B = \{x_0, \ldots, x_n\}$ given by the black dots in the two diagrams above and t.p.m. $P(x \to \cdot)$ which is a symmetric p.m.f. centered at x. If the space B is periodic (top), then x_0 receives support from x_n and x_{n-1}. However when B has edges (bottom), a transition from x_n or x_{n-1} to x_0 is no longer possible. In order 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 $P(x_0 \to \cdot)$ hits x_{-1} or x_{-2}.

2 Examples

2.1 Metropolis

Auxiliary variables lend insight into the Metropolis et al. (1953) algorithm, which can be regarded as an auxiliary variable algorithm itself. As an example we consider a Markov chain defined on the set $B = \{x_0, \ldots, x_n\}$ where the transition probability matrix (t.p.m.) $P(x_i \to \cdot)$ is a symmetric probability mass function centered at x_i with support on $\{x_{i-2}, \ldots, x_{i+2}\}$ 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.

However 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 t.p.m. 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 Figure 1. Here the uniform Metropolis kernel is defined by

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

The term $\sum_{x_k \notin B} P(x_i \to x_k)$ 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, $P(x_i \to x_k) = P(x_k \to x_i)$ which is equal to $P(x_{k \bmod n+1} \to x_i)$ when B is periodic. Hence each term $P(x_i \to x_k)$ with $x_k \notin B$ mimics the term $P(x_{k \bmod n+1} \to x_i)$ 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 proven quite useful in a number of applications in Damien, Wakefield and Walker (1997). If $\pi(x)$ is not readily invertible, an alternative is to use an adaptive approach to sample from $\pi(x|u)$.

The Metropolis algorithm simply replaces step 2 above with edge correction update for uniform distributions P_{MU}. Hence, the resulting updates are

1. $u|x \sim U[0, \pi(x)]$
2. Replace x by x' sampled from $P_{MU}(x \to x')$ with $B_u = \{x : \pi(x) \geq u\}$:
   (a) Sample x^c from the symmetric kernel $P(x \to x^c)$.
   (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 $P(x \to \cdot)$; the chain moves to x^c if x^c is in the set $B_u = \{x : \pi(x) \leq u\}$ or remains at x if x^c falls outside of B_u. After reordering the steps, one recognizes this as the standard Metropolis algorithm for $\pi(x)$, as described in the statistical literature; see Tierney (1994) or Besag, Green, Higdon and Mengersen (1995), for example.

2.2 Swendsen-Wang

The advent of the Swendsen-Wang (SW) algorithm for Ising and Potts models has lead 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 its success with Ising and Potts models, much effort has been spent on generalizing the Swendsen Wang algorithm to a wider class of models. Examples include continuous spin models (Wolff, 1989; De Meo and Oh, 1992) and gray level imaging (Green, 1992). Below we describe the general form of the Swendsen-Wang algorithm due to Edwards and Sokal (1988).

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

$$\pi(x) \propto \pi_0(x) \prod_k b_k(x), \tag{2}$$

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

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

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

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

Now $x|u$ is distributed according to $\pi_0(x)$, subject to the constraints, $u_k \leq b_k(x) \forall k$. In Bayesian image analysis, $\pi_0(x)$ will typically denote 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 example below. However, in general, the constraints can make simulating from $\pi(x|u)$ a difficult task.

Swendsen-Wang for binary Markov random fields

Let $\mathbf{S}$ denote the set of pixels or lattice sites, with pairwise adjacencies indicted 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

$$\pi(x) \propto \exp\left\{\sum_{i\in\mathbf{S}} \alpha_i(x_i)\right\} \times \exp\left\{\sum_{i\sim j} \beta_{ij} I[x_i = x_j]\right\}, \; x \in \{0,1\}^n. \tag{5}$$

Note that the Potts (1952) model has identical form except that each x_i may take on more than two unordered states. We assume that $\beta_{ij} > 0$, so this distribution invites clustering of like colored pixels. In the case of no external field ($\alpha_i(\cdot) \equiv 0 \; \forall i$), we 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 x, the components u_{ij} are then specified to be independent and uniformly distributed

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

so that

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

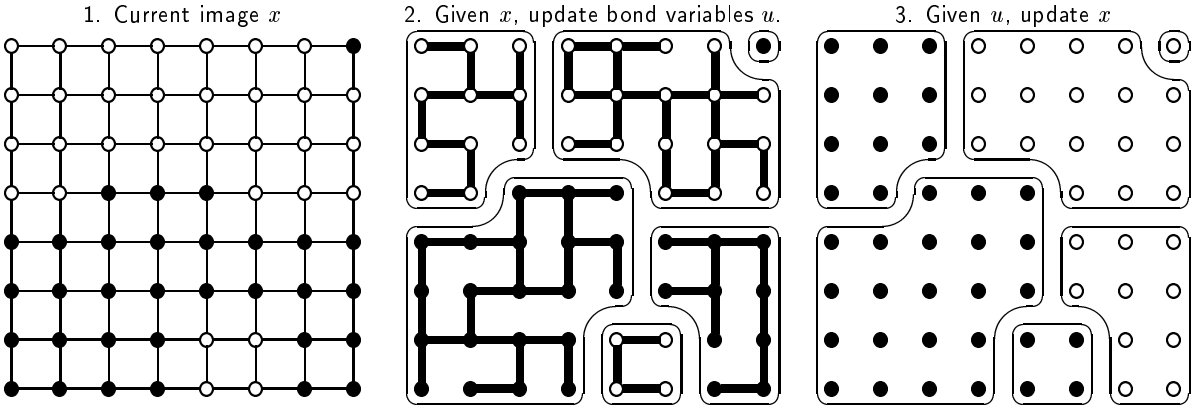
If $u_{ij} > 1$ then $\exp\{\beta_{ij} I[x_i = x_j]\} > 1$, so the condition $u_{ij} > 1$ implies $x_i = x_j$. Like colored neighbors, i and j, are bonded (ie. constrained to be equal) with probability $1 - \exp\{-\beta_{ij}\}$. The bond variable u partitions $\mathbf{S}$ into like colored clusters. For a particular cluster, $\mathbf{C}$, the probability of color $k \in \{0,1\}$ is $\propto \exp\{\sum_{i\in\mathbf{C}} \alpha_i(k)\}$, so each cluster can be updated independently according to its conditional distribution. The bond variables completely decouple the external field term, $\exp\{\sum_{i\in\mathbf{S}} \alpha_i(x_i)\}$, 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 Swendsen-Wang algorithm on an 8×8 lattice with first order neighborhood structure and no external field. Figure 3 shows successive realizations of the Swendsen-Wang algorithm and Gibbs sampling on a 100×100 lattice with all βij set to the critical value β^* for the infinite lattice, $\beta^* \approx 0.88$.

Swendsen-Wang for a gray-level model

For the Ising model, SW gives the most improvement of single site Metropolis when β 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. Though the commonly used Gaussian pairwise difference prior (Besag et al. 1995, Sec 3) doesn't exhibit this

An illustration of the Swendsen-Wang algorithm for the Ising model on the 8×8 lattice.

$$\pi(x) \propto \exp\left\{\beta \sum_{i \sim j} I[x_i = x_j]\right\}$$



1. Initial image x and Markov random field graph for $\pi(x)$.

2. 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.

3. Given the bond variables u, x is now a coarse image of independent clusters. Since there is no external field in this example, each cluster is recolored black or white with probability 0.5.

Figure 2: The Swendsen-Wang algorithm

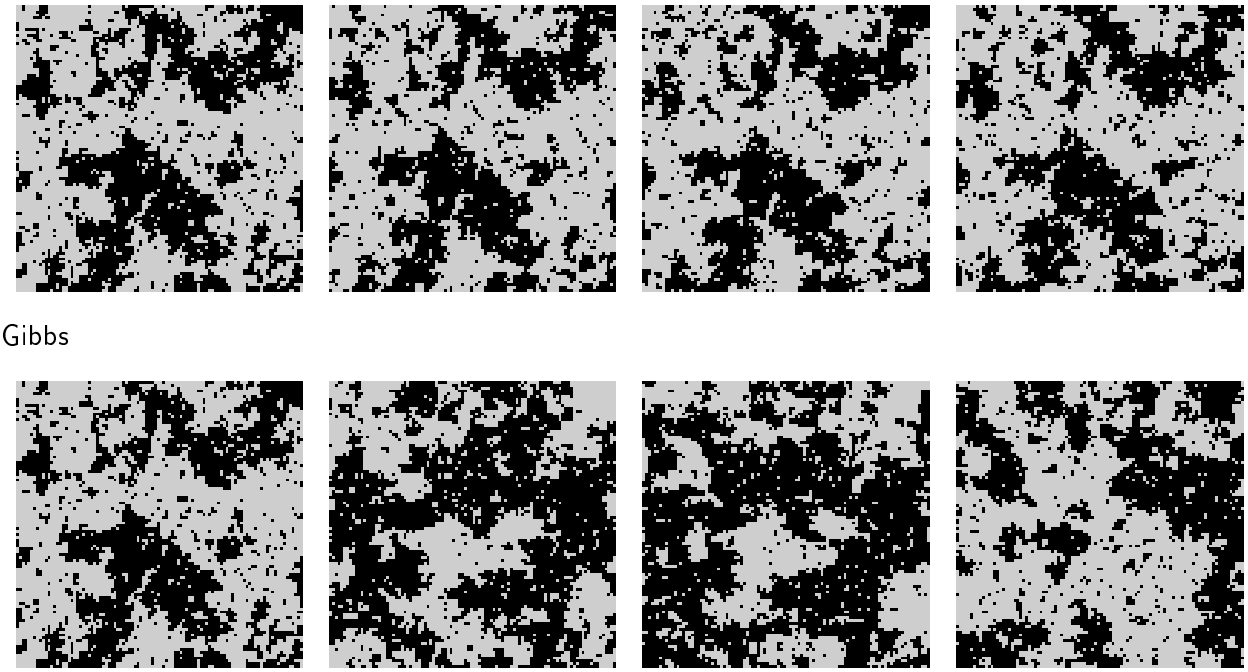


Figure 3: Successive realizations of the Ising model at critical temperature from the single-site Metropolis and Swendsen-Wang algorithms.

behavior, the prior model of Geman and McClure (1987)

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

does show some similarities to the Ising model which suggests 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 $(\beta,\lambda,M) = (0.96,.005,300)$. These parameter values result 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:

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

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

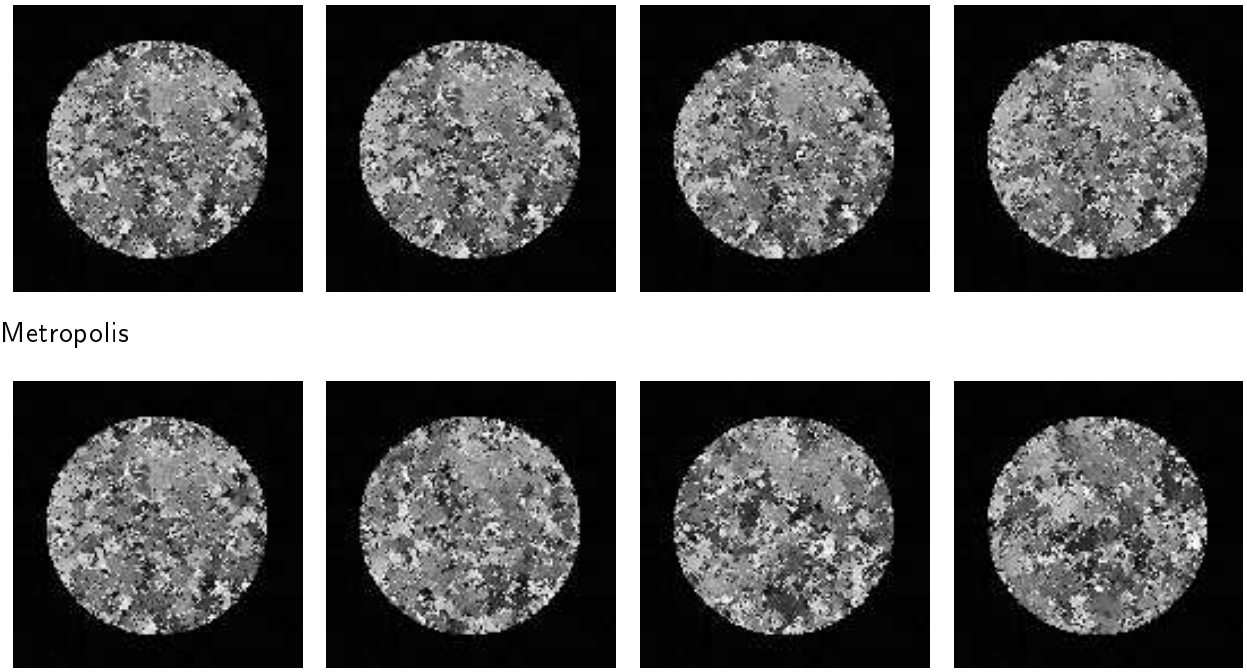


Figure 4: Successive realizations of the Geman and McClure's prior at "critical temperature" from the single-site Metropolis and the generalized Swendsen-Wang algorithms.

2. Determine clusters $x_{\mathbf{C}_1}, \ldots, x_{\mathbf{C}_n}$ induced by bond variable u.

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

Though the above algorithm results in an irreducible Markov chain, we 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. Since 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.

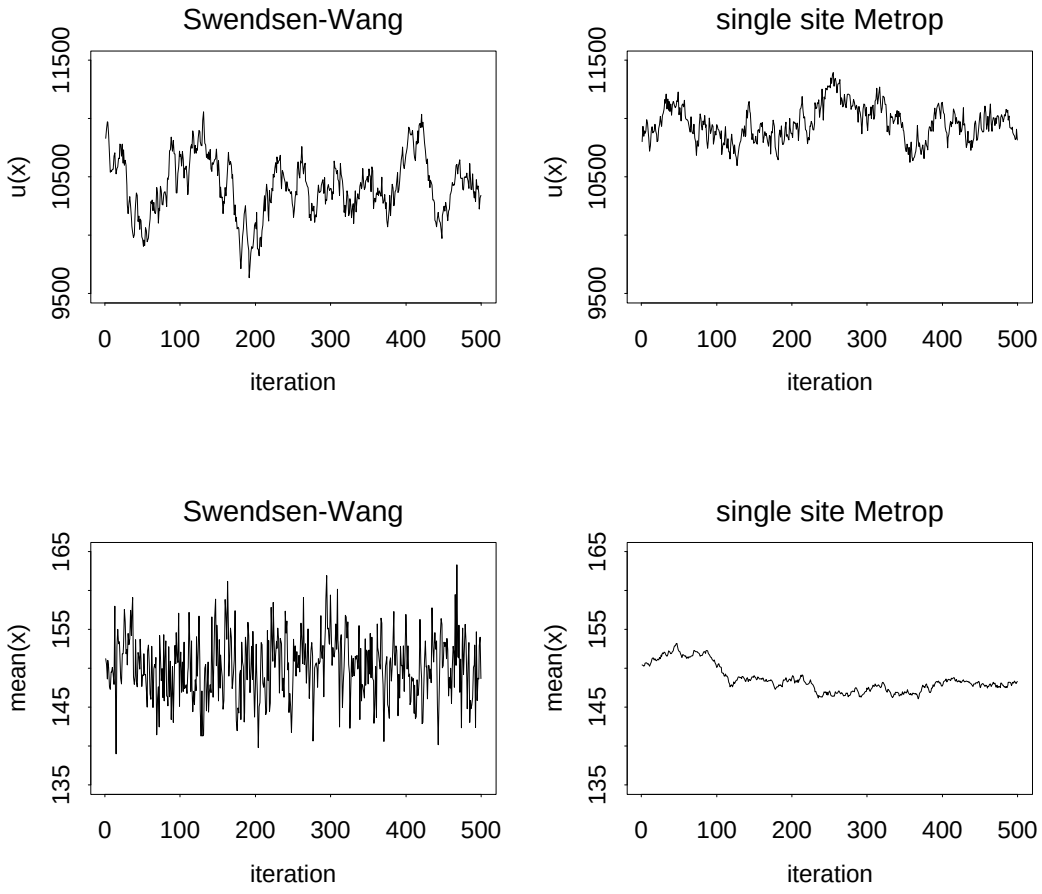


Figure 5: Comparison of time series plots of $u(x) = \sum_{i \sim j}[1 + \lambda(x_i - x_j)^2]^{-1}$ and $\sum x/n$ from MCMC simulation of the Geman-McClure prior using Swendsen-Wang and single site Metropolis updating.

2.3 Partial-decoupling

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

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

leading to the conditional distribution of $x|u$

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

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 occurrs in SW. In partial decoupling, $\pi(u|x)$ is modified so that the prior is only partially decoupled from the likelihood term when considering $\pi(x|u)$. Below we give the details of applying partial decoupling to binary Markov random fields.

Partial decoupling for binary Markov random fields

The method of partial decoupling was originally developed for binary image applications which result in a posterior distribution of the form (5). We set

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

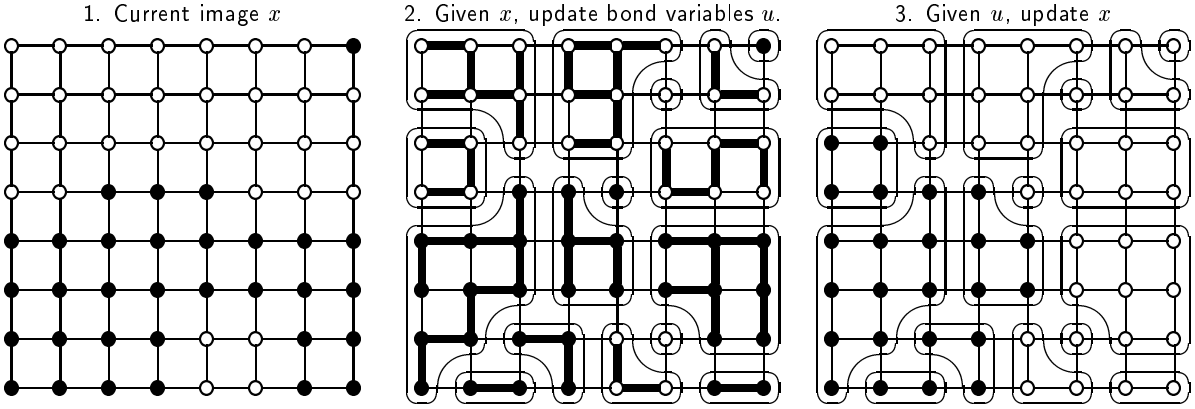
which yields

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

The term $\prod_{i \sim j} I[0 \le u_{ij} \le \exp\{\delta_{ij}\beta_{ij} I[x_i = x_j]\}]$ 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 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 Gibbs, Metropolis, or an auxiliary variable method for this coarsened model. The bonding probability is controlled by the δ_{ij}'s; note that all $\delta_{ij} = 0$ corresponds to Gibbs sampling, while all $\delta_{ij} = 1$ corresponds

An illustration of the partial-decoupling algorithm, with $0 < \delta_{ij} = \delta < 1$ for the Ising model on the 8×8 lattice.

$$\pi(x) \propto \exp\left\{\beta \sum_{i \sim j} I[x_i = x_j]\right\}$$



1. Initial image x and Markov random field graph for $\pi(x)$.

2. 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; though the auxiliary vector u may still lead to constraints, the dependence between neighboring sites is never completely removed.

3. Since 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.

Figure 6: The partial decoupling algorithm

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 8×8 lattice with first order neighborhood structure.

Choosing δ

In practice one is left to specify the constants δ_{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, a number of strategies may be employed. If certain components of x are fixed, or nearly fixed by the likelihood, SW updating can fare very poorly. The problem arises because most of the sites belong to a cluster which contains one or more of the fixed sites. Since the cluster contains the fixed site, the probability of the cluster changing color is zero. Hence most of the image remains unchanged after each iteration. If x_i is fixed at 0 for instance, setting $\delta_{ij} = 0 \; \forall j \in \partial i$, and leaving the remaining δ's at 1 will keep any neighboring site from bonding with the fixed site, while using

the SW scheme away from the fixed site. For an example, see (Higdon, 1994).

Another setting in which the SW algorithm can lead to slow mixing is when the external field term of the Ising model results 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 which form may have very little chance of changing in the presence of the external field. Perhaps the simplest approach is to set $\delta_{ij} = \delta$. Since δ controls the chance neighboring pixels will bond, setting δ to a sufficiently small constant will ensure clusters do not grow too large. An alternative is to choose the δ_{ij}'s to prevent clusters from growing too large or across certain boundaries. For example, the lattice may be broken into sub-lattices. By setting $\delta_{ij} = 0$ for adjacencies linking sub-lattices, while leaving the remaining δ's at 1, clusters cannot grow 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 δ each iteration according to some deterministic schedule. See Hurn (1997) for an application which adaptively blocks the array into successively coarser sub-lattices.

Perhaps the most appealing method for determining δ 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, we choose δ_{ij} to be near 1 if y_i and y_j are similar; we choose δ_{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 we 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 in 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 first is binary classification problem where 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 is a PET (positron emission tomography) 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

As an example 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, we take a Bayesian imaging approach; we base the likelihood on the original image y and assign 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 the left frame of Figure 7. The resulting posterior distribution is given below

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

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 8 neighbors.

The parameters (α, β) are fixed at $(1, .8)$. The value for the Ising parameter β is specified so that posterior realizations x will contain very little speckle, while the value for α 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. Though 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 7 shows the estimated posterior mean for x from a partial decoupling run of 40,000 iterations.

A number of partial decoupling schemes were tried 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, we focus on a subset of the image marked by the 20×20 square outlined in the right frame of Figure 7. This square S is one of the regions which exhibits multimodality. Some realizations show a sizable patch of about 50 – 80 pixels; other realizations do not. For various samplers, we monitor the number of transitions between the two modes. The results are summarized in Table 1. For this particular example, a value of $a = 0.6$ is close to optimal and vastly superior to single-site Metropolis, SW, or even

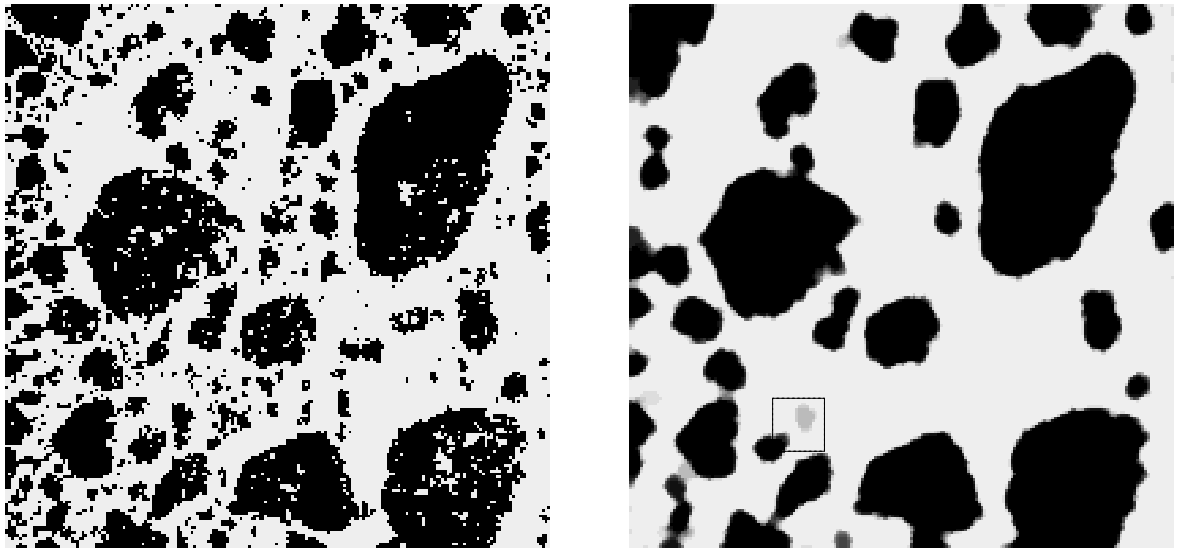


Figure 7: Left frame: Data records y. Right frame: Posterior mean estimate of x from a partial decoupling run of 40,000 iterations. The box marks a 20×20 pixel region which exhibits multimodality.

blocked SW ($a = 1$).

Table 1: Mean number of iterations between mode swaps

Metrop	Partial decoupling, $\delta_{ij} = aI[y_i = y_j]$							SW
$a = 0$	$a = .4$	$a = .5$	$a = .6$	$a = .7$	$a = .8$	$a = .9$	$a = 1$	
20000	2778	758	529	680	741	1818	3448	∞

3.2 A PET application

This final application comes from medical imaging using positron emission tomography (PET). Here one constructs an emission intensity map of an object using photon counts detected by a gamma camera ring, or photon detector ring. This application considers reconstructing a two-dimensional slice of a 3-d object. Figure 8 gives a diagram of the information obtained during a PET scan. As a positron is emitted, it is immediately annihilated by an electron, causing two photons to be emitted in directly opposite directions. When the detector ring registers simultaneous "hits", this defines a thin column or slice through the object which must contain the emission source, pixel i. Columns are indexed by their angle a and bin b as shown in Figure 8. The total number of simultaneous

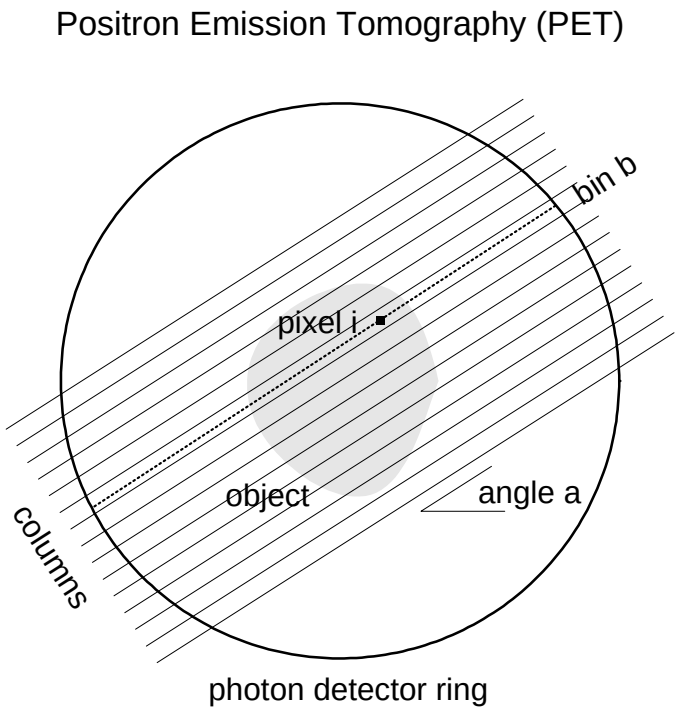


Figure 8: PET reconstructions: A pixelated object emits positrons with location dependent intensities x_i. Immediately after a positron is emitted from pixel i an annihilation event causes two photons to be emitted in directly opposite directions so that pixel i must be contained within a thin column determined by the photon paths. Each pair of "hits" corresponds to a column which is indexed by angle of the column a and the bin number b, giving binned counts y_{ab}. Since 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.

"hits" for each angle and bin are collected. This example uses 512 angles and 96 bins to index the possible columns.

The data consist of counts y_{ab} obtained from the column corresponding to angle a and bin b. Physical characteristics of the object and imaging system determine the chance of a photon pair emanating from pixel i registers on a column corresponding to bin b and angle a. For this application the probabilities p_{abi} were determined through previous experimentation with the imaging system. Thus the counts y_{ab} have a Poisson distribution with mean

$$\mu_{ab} = \sum_i x_i p_{abi}.$$

Given the object intensities x_i, the likelihood can be written

$$L(y|x) \propto \prod_{ab} \mu_{ab}^{y_{ab}} e^{-\mu_{ab}}.$$

A more detailed derivation of the likelihood for emission computed tomography is given in Shepp and Vardi (1982).

We use the prior distribution of Geman and McClure (1987) given in (8) to induce prior structure on the source intensities x. Rather than condition on a single value of β, a hierarchical prior is specified for the interaction parameter so that it may vary throughout the MCMC simulation. To match the previous example in Section 2.2, the pixel intensities are restricted to lie between 0 and $M = 300$, and the parameter λ is held fixed at 0.005. Of course one could rescale the distribution for x without changing $Z(\beta)$ by setting $M = 300 \cdot \sigma$ and $\lambda = .005/\sigma^2$ for arbitrary $\sigma > 0$.

From simulation experiments, if λ and M are fixed at the above values, then the "critical" value β^* is near 0.95. Since we expect the posterior distribution for β to be near β^*, we specify $\pi(\beta)$ to be uniform over S which contains equally spaced values for β between 0.8 and 1.2. 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). Alternativly, thermodynamic integration (Ogata and Tanemura, 1984) or path sampling (Gelman and Meng, 1996) could have been employed. 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

$$\pi(x, \beta|y) \propto \prod_{ab} \mu_{ab}^{y_{ab}} e^{-\mu_{ab}} \times \frac{1}{Z(\beta)} \exp\left\{ \beta \sum_{i \sim j} \left[1 + \lambda(x_i - x_j)^2\right]^{-1} \right\}, \; x \in [0, M]^n, \; \beta \in S.$$

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 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 β is updated via a Metropolis step which 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 [.96,.97]. Of course this is only an initial attempt to treat β as a parameter in the MCMC simulation. A more thorough analysis should also consider the choices of λ and M, possibly treating them as dynamic parameters in the simulation.

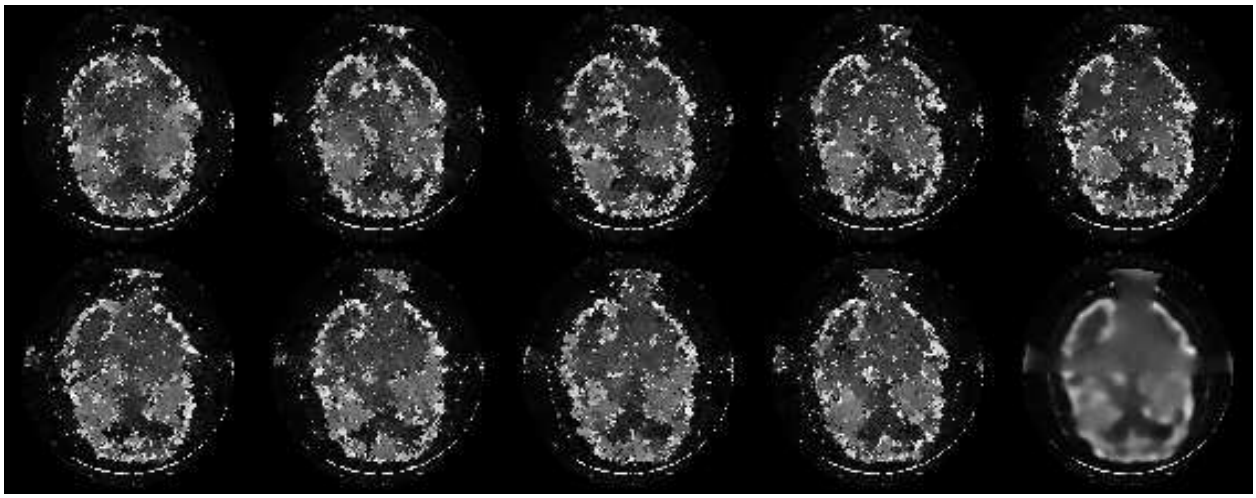


Figure 9: Realizations from the posterior generated using SW updates for the pixel values x and Metropolis for β. The final image is an estimate of the posterior mean for x.

4 Discussion

The applications presented here show that auxiliary variable methods can lead to substantial gains in efficiency when using MCMC 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 where such algorithms are likely to fare better than standard single site approaches. Though there are no clear rules yet 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 (Section 2.2) have similarities that are worth noting. They both respond well to SW updating, but not 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 will likely be quite successful.

The two applications in Section 3 differ in the relative strengths of their prior and likelihood. 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's very unlikely

a cluster will form that can change color. Partial decoupling is effective here because it governs the cluster size and forms clusters which 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 since 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 non-conjugate formulations in 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) show the slice sampler is superior to an independence Metoroplis sampler. This isn't surprising in light of the reexpression of the Metropolis algorithm as an auxiliary variable method in Section 2.1. Though some may be unhappy with the increase in dimensionality for slice samplers, this increase also occurs implicitly for Metropolis samplers as well. In addition, the slice sampler uses a Gibbs update for $x|u$ where Metropolis uses the clunky edge correction kernel.

We finish 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 paper, 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.

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, Dept of Operations Research, University of North Carolina.

Gelman, A., and Meng, X. (1996), "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling," Technical report 440, Department of Statistics, University of Chicago.

Geman, S., and McClure, D. (1987), "Statistical methods for tomographic image reconstruction," *Bulliten of the International Statistical Institute*, 52, no.4 5–21.

Geyer, C. J. (1991), "Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo," Technical Report 568, School of Statistics, University of Minnesota.

Geyer, C. J. (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, Statistics Group, University of Bristol.

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.

Higdon, D. M. (1994), "Spatial applications of Markov chain Monte Carlo for Bayesian inference," unpublished Ph.D. dissertation, Department of Statistics, University of Washington.

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," *Phisical Review, 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, School of Statistics, University of Minnesota.

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, Statistical Laboratory, University of Cambridge.

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 Troisiéme 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), *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.