Bayesian Variable Selection and the Swendsen-Wang Algorithm

David J. NOTT and Peter J. GREEN

The need to explore model uncertainty in linear regression models with many predictors has motivated improvements in Markov chain Monte Carlo sampling algorithms for Bayesian variable selection. Currently used sampling algorithms for Bayesian variable selection may perform poorly when there are severe multicollinearities among the predictors. This article describes a new sampling method based on an analogy with the Swendsen-Wang algorithm for the Ising model, and which can give substantial improvements over alternative sampling schemes in the presence of multicollinearity. In linear regression with a given set of potential predictors we can index possible models by a binary parameter vector that indicates which of the predictors are included or excluded. By thinking of the posterior distribution of this parameter as a binary spatial field, we can use auxiliary variable methods inspired by the Swendsen-Wang algorithm for the Ising model to sample from the posterior where dependence among parameters is reduced by conditioning on auxiliary variables. Performance of the method is described for both simulated and real data.

Key Words: Auxiliary variables; Data augmentation; Ising model; Markov chain Monte Carlo.

1. INTRODUCTION

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be a vector of responses, $\mathbf{X}$ be an $n \times p$ design matrix, and consider a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of parameters and $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$ is a vector of zero mean errors. We consider Bayesian inference in this model with a hierarchical prior on $\boldsymbol{\beta}$ which allows some components of $\boldsymbol{\beta}$ to be zero. If $\beta_i = 0$, this excludes the ith predictor from the model. The problem of variable selection is to decide which predictors should

David J. Nott is Lecturer, Department of Statistics, University of New South Wales, Sydney 2052, Australia (E-mail: djn@maths.unsw.edu.au). Peter J. Green is Professor, Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK (E-mail: P.J.Green@bristol.ac.uk).

Journal of Computational and Graphical Statistics, Volume 13, Number 1, Pages 1–17
DOI: 10.1198/1061860042958

be included in a model for the mean of the responses; the Bayesian approach does this coherently, integrating out all uncertainties.

Markov chain Monte Carlo sampling methods for exploring model uncertainty in Bayesian variable selection problems have received a lot of recent attention: see George and McCulloch (1997), Denison, Mallick, and Smith (1998) and Kohn, Smith, and Chan (2001) for a discussion of different approaches and recent developments.

Currently used sampling schemes for exploring the posterior distribution may mix slowly when there are severe multicollinearities among the predictors. We describe an algorithm which offers improvements over alternative sampling schemes in this situation, and which is based on an analogy with the Swendsen-Wang algorithm for the Ising model (Swendsen and Wang 1987). We formulate our hierarchical prior for $\boldsymbol{\beta}$ in terms of a vector of binary variables in which the components indicate whether a predictor is included in the model or not. By thinking of this binary parameter vector as a spatial process, we are motivated to use a sampling algorithm inspired by the Swendsen-Wang algorithm for the Ising model, where dependence between parameters is reduced by conditioning on some auxiliary variables. For a review of the Swendsen-Wang algorithm and some extensions to general Bayesian inference see Higdon (1998).

We will be concerned with Bayesian methods for variable selection and accounting for model uncertainty in linear models. However, similar ideas find application in many other areas, such as generalized linear models (Raftery 1996), survival analysis (Volinsky, Madigan, Raftery, and Kronmal 1997) and graphical models (Madigan and York 1995).

The structure of this article is as follows. Section 2 specifies the model and the priors. Section 3 reviews the Swendsen-Wang algorithm, and Section 4 extends the algorithm to the problem of Bayesian variable selection. Section 5 describes our method for defining the auxiliary variables in our algorithm. Section 6 describes performance of the method for real and simulated data, and Section 7 gives some discussion and conclusions.

2. BAYESIAN VARIABLE SELECTION

We consider the model and the prior specification given by Kohn et al. (2001) for Bayesian variable selection problems, although the methods we describe are applicable with other prior specifications. Following their notation, let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ be a binary vector, and write $q_\gamma = \sum_i \gamma_i$ for the number of nonzero elements of $\boldsymbol{\gamma}$. Let $\mathbf{X}_\gamma$ be the $n \times q_\gamma$ design matrix obtained by removing those columns i from $\mathbf{X}$ for which $\gamma_i = 0$. Similarly let $\boldsymbol{\beta}_\gamma$ be the subvector of $\boldsymbol{\beta}$ obtained by removing components β_i of $\boldsymbol{\beta}$ for which $\gamma_i = 0$.

We assume that

$$\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}_\gamma, \boldsymbol{\beta}_\gamma, \sigma^2 \sim N(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \mathbf{I}).$$

For Bayesian inference on the model parameters we use a hierarchical prior. The prior for $\boldsymbol{\beta}_\gamma$ given $\boldsymbol{\gamma}$ and σ^2 is normal

$$p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2) \quad \sim \quad N(\mathbf{0}, n\sigma^2(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}). \tag{2.1}$$

This prior was used by Smith and Kohn (1996), is related to the g-prior of Zellner (1986), and has some attractive invariance properties under rescaling of X and y (see Kohn et al. 2001 for further discussion).

The prior on σ^2 is $p(\sigma^2) \propto \sigma^{-2}$, and for our prior on $\boldsymbol{\gamma}$ we use

$$p(\boldsymbol{\gamma}|\pi) = \prod_{i=1}^{p} \pi^{\gamma_i}(1-\pi)^{1-\gamma_i} = \pi^{q_\gamma}(1-\pi)^{p-q_\gamma},$$

where π is a hyperparameter with a beta prior, $\pi \sim \text{Beta}(a, b)$. The prior on $\boldsymbol{\gamma}$ (integrating out π) is

$$p(\boldsymbol{\gamma}) = \frac{B(q_\gamma + a, p - q_\gamma + b)}{B(a, b)},$$

where $B(\cdot, \cdot)$ denotes the beta function.

We are interested in the posterior distribution on $\boldsymbol{\gamma}$ with $\boldsymbol{\beta}_\gamma$ and σ^2 integrated out,

$$p(\boldsymbol{\gamma}|\mathbf{y}) \quad \propto \quad p(\boldsymbol{\gamma})p(\mathbf{y}|\boldsymbol{\gamma}). \tag{2.2}$$

We have that

$$p(\mathbf{y}|\boldsymbol{\gamma}) = \int\int p(\mathbf{y}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \sigma^2)p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2)p(\sigma^2)d\boldsymbol{\beta}_\gamma d\sigma^2$$

and Smith and Kohn (1996) observed that $\boldsymbol{\beta}_\gamma$ can be integrated out as a normal integral, and σ^2 as an inverse gamma integral, to give

$$p(\mathbf{y}|\boldsymbol{\gamma}) \propto (1+n)^{-q_\gamma/2}\left(\mathbf{y}^T\mathbf{y} - \frac{n}{n+1}\mathbf{y}^T\mathbf{X}_\gamma(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^T\mathbf{y}\right)^{-n/2}.$$

For alternative prior specifications it may not be possible to integrate out $\boldsymbol{\beta}_\gamma$ and σ^2 analytically: if this is the case, then we can still approximate $p(\mathbf{y}|\boldsymbol{\gamma})$ via a Laplace approximation and apply the methods described later. This may also be useful in developing sampling schemes for other Bayesian model selection problems.

For p relatively small, we can compute the posterior $p(\boldsymbol{\gamma}|\mathbf{y})$ exactly, obtaining the normalizing constant in (2.2) by summing over all possible values of $\boldsymbol{\gamma}$. For large p, this is not feasible due to the number of terms in the sum, and we use Markov chain Monte Carlo algorithms to identify high posterior probability models.

When there are high posterior correlations between components of $\boldsymbol{\gamma}$, the usual Markov chain Monte Carlo methods for exploring the posterior, which update one component of $\boldsymbol{\gamma}$ at a time, can mix slowly. High posterior correlations can occur, for instance, in the situation where there is multicollinearity. Updating components of $\boldsymbol{\gamma}$ in blocks rather than one at a time can alleviate problems of slow convergence, but it may be difficult to decide how to choose blocks.

Although the focus of this article is on variable selection, also of interest is estimation of the vector of regression coefficients $\boldsymbol{\beta}$ without conditioning on any single model, but by averaging over different models (so-called model averaging, see Hoeting, Madigan, Raftery, and Volinsky 1999). The sampling schemes we discuss are relevant for this objective also.

3. SWENDSEN-WANG ALGORITHM

Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)$ be a binary spatial process with joint distribution $p(\boldsymbol{\eta})$ specified by

$$p(\boldsymbol{\eta}) \quad \propto \quad \exp\left(\sum_i \alpha_i(\eta_i) + \sum_{i<j} \psi_{ij} I(\eta_i = \eta_j)\right), \tag{3.1}$$

where $\psi_{ij} \geq 0, i < j$, and $I(A)$ is the indicator function which is one when A occurs and zero otherwise. Ratios of the expression (3.1) are easy to compute for $\boldsymbol{\eta}$ vectors which differ at a single site, and this allows a single site Metropolis-Hastings algorithm such as the Gibbs sampler to be easily implemented. However, single-site updating schemes can mix slowly when there is strong dependence between components of $\boldsymbol{\eta}$.

An alternative to the usual single-site updating schemes is the Swendsen-Wang algorithm, in which auxiliary variables are introduced which conditionally remove interactions among components of $\boldsymbol{\eta}$. We let $\mathbf{u} = \{u_{ij} : 1 \leq i < j \leq p\}$ be a set of auxiliary variables and set up a joint distribution $p(\mathbf{u}, \boldsymbol{\eta})$ on $\mathbf{u}$ and $\boldsymbol{\eta}$ in which the marginal distribution for $\boldsymbol{\eta}$ is given by (3.1). This joint distribution can be constructed so that $p(\mathbf{u}|\boldsymbol{\eta})$ and $p(\boldsymbol{\eta}|\mathbf{u})$ are easy to sample from.

To give the joint distribution $p(\mathbf{u}, \boldsymbol{\eta})$ for $\mathbf{u}$ and $\boldsymbol{\eta}$ we specify $p(\mathbf{u}|\boldsymbol{\eta})$. Given $\boldsymbol{\eta}$, the u_{ij} are mutually independent with the distribution of u_{ij} uniform,

$$p(u_{ij}|\boldsymbol{\eta}) = \frac{1}{\exp(\psi_{ij} I(\eta_i = \eta_j))} I(u_{ij} \in [0, \exp(\psi_{ij} I(\eta_i = \eta_j))]).$$

Then we have

$$\begin{aligned} p(\mathbf{u}, \boldsymbol{\eta}) &= p(\boldsymbol{\eta}) p(\mathbf{u}|\boldsymbol{\eta}) \\ &\propto \exp\left\{\sum_i \alpha_i(\eta_i)\right\} I(u_{ij} \in [0, \exp(\psi_{ij} I(\eta_i = \eta_j))] \; \forall i, j) \end{aligned}$$

and, of course, $p(\boldsymbol{\eta}|\mathbf{u}) \propto p(\mathbf{u}, \boldsymbol{\eta})$.

If $\psi_{ij} > 0$, then the condition $u_{ij} < \exp(\psi_{ij} I(\eta_i = \eta_j))$ is satisfied if $u_{ij} < 1$, or even if $u_{ij} \in [1, \exp(\psi_{ij})]$ if $\eta_i = \eta_j$. So when $u_{ij} \in [1, \exp(\psi_{ij})]$, η_i and η_j are constrained to be equal by $p(\boldsymbol{\eta}|\mathbf{u})$, and the auxiliary variables u_{ij} thus define clusters of sites with the same value. Subject to these constraints, we see from the expression for $p(\boldsymbol{\eta}|\mathbf{u})$ that those components of $\boldsymbol{\eta}$ not constrained to be equal are conditionally independent. Conditioning on the auxiliary variables removes the interactions between the components of $\boldsymbol{\eta}$. Hence both $p(\mathbf{u}|\boldsymbol{\eta})$ and $p(\boldsymbol{\eta}|\mathbf{u})$ are easy to sample from.

Higdon (1998) introduced a modification of the ordinary Swendsen-Wang algorithm which he calls partial decoupling. Applied to the Ising model, the idea of partial decoupling is that it may be helpful in the Swendsen-Wang algorithm to define the auxiliary variables as

$$p(u_{ij}|\boldsymbol{\eta}) = \frac{1}{\exp(\psi^*_{ij} I(\eta_i = \eta_j))} I(u_{ij} \in [0, \exp(\psi^*_{ij} I(\eta_i = \eta_j))]),$$

where the ψ_{ij}^* are obtained by scaling down the values ψ_{ij}. The parameters ψ_{ij}^* defining the auxiliary variables are a free choice in the algorithm and they need not be taken the same as the model parameters ψ_{ij}. It may be beneficial to scale down the ψ_{ij} values when defining the auxiliary variables to improve mixing in the algorithm. The reason for this is that if we put $\psi_{ij} = \psi_{ij}^*$, and if many of the ψ_{ij} are large, then in general there will be large clusters of the variables involved in constraints at each step of the sampling algorithm, and often at least one of the η_i in a large cluster will have a value fixed by the likelihood, making it difficult to change cluster values.

4. ANALOGY WITH BAYESIAN VARIABLE SELECTION

By thinking about the posterior distribution (2.2) as a binary spatial process we can construct an MCMC algorithm analogous to the Swendsen-Wang algorithm which performs better than single site updating schemes for exploring the posterior in the presence of strong posterior correlations.

As in the Swendsen-Wang algorithm, let $\mathbf{u} = \{u_{ij} : 1 \leq i < j \leq p\}$ be a collection of auxiliary variables, and define a joint distribution $p(\mathbf{u}, \boldsymbol{\gamma}|\mathbf{y})$ as $p(\mathbf{u}, \boldsymbol{\gamma}|\mathbf{y}) = p(\boldsymbol{\gamma}|y)p(\mathbf{u}|\boldsymbol{\gamma}, \mathbf{y})$ where $p(\boldsymbol{\gamma}|\mathbf{y})$ is the posterior distribution (2.2) and $p(\mathbf{u}|\boldsymbol{\gamma}, \mathbf{y})$ must be chosen. We have a great deal of freedom in how we may choose $p(\mathbf{u}|\boldsymbol{\gamma}, \mathbf{y})$.

We choose $p(\mathbf{u}|\boldsymbol{\gamma}, \mathbf{y})$ as in the Swendsen-Wang algorithm,

$$p(\mathbf{u}|\boldsymbol{\gamma}, \mathbf{y}) = \frac{1}{\exp(\sum_{i<j} \psi_{ij} I(\gamma_i = \gamma_j))} I(u_{ij} \in [0, \exp(\psi_{ij} I(\gamma_i = \gamma_j))] \; \forall i < j),$$

where ψ_{ij} are some (possibly negative) interaction parameters. We discuss how these interaction parameters are determined in the next section.

Then

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{u}, \mathbf{y}) &\propto p(\boldsymbol{\gamma}|\mathbf{y})p(\mathbf{u}|\boldsymbol{\gamma}, \mathbf{y}) \\ &= \frac{p(\boldsymbol{\gamma}|\mathbf{y})}{\exp(\sum_{i<j} \psi_{ij} I(\gamma_i = \gamma_j))} I(u_{ij} \in [0, \exp(\psi_{ij} I(\gamma_i = \gamma_j))] \; \forall i < j). \end{aligned}$$

Although $p(\boldsymbol{\gamma}|\mathbf{y})$ does not in general have the form (3.1), it is hoped that as in the Swendsen-Wang algorithm the denominator in the above expression will serve to reduce interactions among components of $\boldsymbol{\gamma}$ conditional on $\mathbf{u}$ for a suitable choice of the interaction parameters. This is the key idea in our method.

If $\psi_{ij} > 0$, the constraint $u_{ij} < \exp(\psi_{ij} I(\gamma_i = \gamma_j))$ is satisfied when $u_{ij} < 1$, or even if $u_{ij} \in [1, \exp(\psi_{ij})]$ when $\gamma_i = \gamma_j$. So if $\psi_{ij} > 0$, γ_i and γ_j are constrained to be equal if $u_{ij} \in [1, \exp(\psi_{ij})]$. On the other hand, if $\psi_{ij} < 0$ then the constraint $u_{ij} < \exp(\psi_{ij} I(\gamma_i = \gamma_j))$ is satisfied when $u_{ij} < \exp(\psi_{ij})$, or when $u_{ij} \in [\exp(\psi_{ij}), 1]$ if $\gamma_i \neq \gamma_j$. So if $\psi_{ij} < 0$ and $u_{ij} \in [\exp(\psi_{ij}), 1]$, then γ_i and γ_j are constrained to be different.

The auxiliary variables u_{ij} define clusters among the components of $\boldsymbol{\gamma}$ in much the same way as in the Swendsen-Wang algorithm. Components of $\boldsymbol{\gamma}$ within the same cluster are linked by a set of constraints of the form $\gamma_i = \gamma_j$ or $\gamma_i \neq \gamma_j$. We note that the constraints always have at least one feasible solution, since they are created based on the current value for $\boldsymbol{\gamma}$. Let $C = C(\mathbf{u})$ be one cluster defined by the set of auxiliary variables $\mathbf{u}$, and let $\overline{C}$ denote the set of variables not in C. Let $\boldsymbol{\gamma}(C)$ be the subset of $\boldsymbol{\gamma}$ corresponding to the variables in C, and $\boldsymbol{\gamma}(\overline{C})$ denote the remaining components of $\boldsymbol{\gamma}$. Given the constraints, note that there are only two possible values for the vector $\boldsymbol{\gamma}(C)$: from one possible value we can obtain the other by "flipping" the ones to zeros and zeros to ones within the cluster C.

In general, we can update $\boldsymbol{\gamma}(C)$ by a Metropolis-Hastings step. Write $\boldsymbol{\gamma}^{\text{new}}$ for a proposed value of $\boldsymbol{\gamma}$ in which $\boldsymbol{\gamma}^{\text{new}}(C)$ is generated from the proposal distribution $q(\boldsymbol{\gamma}(C)|\boldsymbol{\gamma}, \mathbf{u}, \mathbf{y})$ and $\boldsymbol{\gamma}^{\text{new}}(\overline{C}) = \boldsymbol{\gamma}(\overline{C})$. The Metropolis-Hastings acceptance probability is

$$\min\left\{1, \frac{q(\boldsymbol{\gamma}(C)|\boldsymbol{\gamma}, \mathbf{u}, \mathbf{y})p(\boldsymbol{\gamma}^{\text{new}}|\mathbf{y})}{q(\boldsymbol{\gamma}^{\text{new}}(C)|\boldsymbol{\gamma}, \mathbf{u}, \mathbf{y})p(\boldsymbol{\gamma}|\mathbf{y})} \exp\left(\sum_{i<j} \psi_{ij}(I(\gamma_i = \gamma_j) - I(\gamma_i^{\text{new}} = \gamma_j^{\text{new}})\right)\right\}.$$

Because $\boldsymbol{\gamma}^{\text{new}}(\overline{C}) = \boldsymbol{\gamma}(\overline{C})$ we can simplify this expression by noting that

$$\exp\left(\sum_{i<j} \psi_{ij}(I(\gamma_i = \gamma_j) - I(\gamma_i^{\text{new}} = \gamma_j^{\text{new}}))\right)$$
$$= \exp\left(\sum_{(i,j)\in\partial C} \psi_{ij}(I(\gamma_i = \gamma_j) - I(\gamma_i^{\text{new}} = \gamma_j^{\text{new}}))\right),$$

where

$$\partial C = \{(i,j) : i < j \quad \text{and either} \quad i \in C, j \notin C \quad \text{or} \quad i \notin C, j \in C\}.$$

This probability is inexpensive to compute provided that ψ_{ij} is nonzero only for a fairly small number of pairs $(i,j) \in \partial C$.

A special case of the general Metropolis-Hastings scheme is to take the cluster proposal to be a Gibbs type proposal, namely the conditional distribution for $\boldsymbol{\gamma}(C)|\boldsymbol{\gamma}(\overline{C}), \mathbf{u}, \mathbf{y}$, which makes the Metropolis-Hastings acceptance probability equal to one. If we set $\psi_{ij} = 0$ for all i, j in this case, then we obtain the Gibbs sampler.

There is also an antithetic method for updating clusters which proposes a change to the current state for a cluster more often than the Gibbs type proposal. Let

$$q = \Pr(\boldsymbol{\gamma}(C)|\boldsymbol{\gamma}(\overline{C}), u, y),$$

be the probability that $\boldsymbol{\gamma}(C)$ remains at its current value (given the current values for $\boldsymbol{\gamma}(\overline{C})$ and $\mathbf{u}$) with the Gibbs type proposal. Now, suppose that instead of using the Gibbs proposal we flip to the opposite state for $\boldsymbol{\gamma}(C)$ rather than remaining with the current with probability

$$\min\left(1, \frac{1-q}{q}\right). \tag{4.1}$$

It is easy to show that in this case detailed balance is maintained.

There are theoretical reasons for always preferring the antithetic method for updating clusters. With the antithetic approach the transition matrix of the chain has uniformly smaller off diagonal entries which results in smaller asymptotic variance of ergodic averages (Peskun 1973, theorem 2.1.1). We consider only the antithetic approach in what follows.

5. OBTAINING THE INTERACTION PARAMETERS

To implement our algorithm for Bayesian variable selection, we need to specify the interaction parameters ψ_{ij} in $p(\mathbf{u}|\boldsymbol{\gamma}, \mathbf{y})$. For computational reasons it is advisable to keep the number of nonzero ψ_{ij} as small as possible (so that the Metropolis-Hastings acceptance probabilities are inexpensive to compute). Also, the nonzero ψ_{ij} should not be too large in magnitude, for the same reason that motivated Higdon's partial decoupling modification of the original Swendsen-Wang algorithm.

We first describe our specification of the interaction parameters and then discuss the motivation for our choice. Let $\boldsymbol{\gamma}^*$ be some fixed configuration for $\boldsymbol{\gamma}$ and define

$$\psi_{ij}^U = 0.5 \times \left(\sum_{\gamma_i=\gamma_j, \gamma_k=\gamma_k^*, k\neq i,j} \log p(\mathbf{y}|\boldsymbol{\gamma}) - \sum_{\gamma_i\neq\gamma_j, \gamma_k=\gamma_k^*, k\neq i,j} \log p(\mathbf{y}|\boldsymbol{\gamma}) \right), \quad (5.1)$$

where $p(\mathbf{y}|\boldsymbol{\gamma})$ is the marginal likelihood of $\boldsymbol{\gamma}$. Then let $\psi_{ij}^S = c\psi_{ij}^U$, where c is a scaling factor chosen so that $\psi_{kl} \in [-a, a]$ for all $k < l$, and finally set $\psi_{ij} = \psi_{ij}^S I(|\psi_{ij}^S| \geq t)$, where t is a truncation point and $I(A)$ is the indicator function which is one when A occurs and zero otherwise. So we have $\psi_{ij} = \psi_{ij}^S$ if $|\psi_{ij}^S| \geq t$ and $\psi_{ij} = 0$ otherwise. The parameters $\boldsymbol{\gamma}^*$, a and t must be chosen to completely specify our approach to choosing interaction parameters. In our examples we use $\boldsymbol{\gamma}^*$ a vector of ones, $a = 1$ and $t = 0.1$.

We now describe the motivation for our choice of ψ_{ij}. First, suppose that $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$ is a binary spatial process with joint distribution of the form (3.1). Let $\boldsymbol{\eta}^* = (\eta_1^*, \ldots, \eta_p^*)^T$ be some fixed configuration for the sites. Then the following formula holds, regardless of the value chosen for $\boldsymbol{\eta}^*$:

$$\psi_{ij} = 0.5 \times \left(\sum_{\eta_i=\eta_j,\, \eta_k=\eta_k^*, k\neq i,j} \log p(\boldsymbol{\eta}) - \sum_{\eta_i\neq\eta_j,\, \eta_k=\eta_k^*, k\neq i,j} \log p(\boldsymbol{\eta}) \right). \quad (5.2)$$

This is easily seen by direct substitution from Equation (3.1). Note that this formula can still be applied even when $\log p(\boldsymbol{\eta})$ is only known up to an additive constant. By analogy with the Swendsen-Wang algorithm, this encourages us to base the choice of ψ_{ij} in our sampling method on Equation (5.2) with $p(\boldsymbol{\eta})$ replaced by $p(\boldsymbol{\gamma}|\mathbf{y})$. Replacing $p(\boldsymbol{\gamma}|\mathbf{y})$ with the marginal likelihood $p(\mathbf{y}|\boldsymbol{\gamma})$ gives Equation (5.1). We discuss the reason for using the marginal likelihood rather than the posterior distribution later. Scaling down the interaction parameters from Equation (5.1) and setting to zero values smaller in magnitude than t gives our method for choosing our algorithm parameters ψ_{ij}.

We note that the value obtained for ψ_{ij}^U in (5.1) will in general depend on the value $\boldsymbol{\gamma}^*$. As mentioned earlier, we use $\boldsymbol{\gamma}^*$ a vector of ones, $a = 1$ and $t = 0.1$. We have experimented with values of a in the range 0.5 to 2.5 and with different values of t, but performance of the algorithm does not seem to be too sensitive to any reasonable choice for these parameters. Choice of $\boldsymbol{\gamma}^*$ is crucial, however. Intuitively, setting all components of $\boldsymbol{\gamma}^*$ to 1 allows us to capture the conditional relationship between γ_i and γ_j in a model in which all important predictors are included. Also, we have found it beneficial to replace use $p(\mathbf{y}|\boldsymbol{\gamma})$ rather than $p(\boldsymbol{\gamma}|\mathbf{y})$ in (5.1). Our reasoning for this is as follows.

Note that we can write $p(\gamma_i, \gamma_j|\mathbf{y}, \boldsymbol{\gamma}_{\neq i,j}) \propto p(\gamma_i, \gamma_j|\boldsymbol{\gamma}_{\neq i,j})p(\mathbf{y}|\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}_{\neq i,j} = \{\gamma_k : k \neq i, j\}$ so that we can rewrite (5.1) as

$$\begin{aligned}\psi_{ij}^U &= 0.5 \times \left(\sum_{\gamma_i=\gamma_j} \log p(\gamma_i, \gamma_j|\boldsymbol{\gamma}_{\neq i,j} = \boldsymbol{\gamma}^*_{\neq i,j}) - \sum_{\gamma_i\neq\gamma_j} \log p(\gamma_i, \gamma_j|\boldsymbol{\gamma}_{\neq i,j} = \boldsymbol{\gamma}^*_{\neq i,j}) \right) \\ &\quad + 0.5 \times \left(\sum_{\gamma_i=\gamma_j,\, \boldsymbol{\gamma}_{\neq i,j}=\boldsymbol{\gamma}^*_{\neq i,j}} \log p(\mathbf{y}|\boldsymbol{\gamma}) - \sum_{\gamma_i\neq\gamma_j,\, \boldsymbol{\gamma}_{\neq i,j}=\boldsymbol{\gamma}^*_{\neq i,j}} \log p(\mathbf{y}|\boldsymbol{\gamma}) \right) \qquad (5.3)\end{aligned}$$

so we can separate out prior and likelihood contributions to ψ_{ij}^U. If we set $\boldsymbol{\gamma}^*_{\neq i,j}$ to a vector of ones, we have found that the relative contribution of the prior in the equation above is not typical of that for models with appreciable posterior probability, since these models may be much more parsimonious than the full model and hence it may be beneficial when using priors which encourage model parsimony to ignore the prior contribution in (5.3) and base calculation of ψ_{ij}^U only on the likelihood $p(\mathbf{y}|\gamma)$ with $\boldsymbol{\gamma}^*_{\neq i,j}$ a vector of ones.

We investigate in the empirical studies of the next section three basic rules for determining the interaction parameters. Our first rule is to simply set $\psi_{ij} = 0$ for all pairs of variables, hereafter referred to as method A. We use this as our baseline method for comparison rather than the Gibbs sampler, since as mentioned in Section 3, this sampler is provably better.

Our second rule for choosing interaction parameters, hereafter method B, chooses ψ_{ij} in the way described above with $\boldsymbol{\gamma}^*$ a vector of ones, $a = 1$ and $t = 0.1$.

For our third rule, method C, we follow method B with the exception of allowing only a much smaller number of the interaction parameters to be nonzero. Computing ψ_{ij}^U for all $i < j$ when there is a large number of predictors in method B can be very time consuming. So if we can reduce the number of pairs i, j for which we must compute the interaction parameter ψ_{ij} then this can improve computational efficiency.

Our method for choosing which pairs i, j have an interaction parameter $\psi_{ij} \neq 0$ involves the use of a standard multicollinearity diagnostic, the variance proportions. For further background see, for instance, Myers (1990). We try to identify severe linear dependencies among columns of the design matrix using the variance proportions, and only allow $\psi_{ij} \neq 0$ for those columns i, j involved in a severe linear dependence.

We now describe the method more precisely. Let $\mathbf{Z}$ be the $n \times p$ matrix obtained from the design matrix $\mathbf{X}$ by centering and scaling each of the predictors (so that each column of $\mathbf{Z}$ is a vector of length one with entries which have zero mean). If an intercept term is fitted, then the corresponding column in the design matrix is scaled to have length one but is not centered. Write $\mathbf{Z}^T\mathbf{Z} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ for the eigenvalue decomposition of $\mathbf{Z}^T\mathbf{Z}$, where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ is the orthogonal matrix with columns given by the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$ of $\mathbf{Z}^T\mathbf{Z}$ and $\mathbf{D}$ is the diagonal matrix of eigenvalues with diagonal entries $\lambda_1, \ldots, \lambda_p$. The least squares estimate $\mathbf{b}$ of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ in the model $\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$ has covariance matrix $\sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$. We can write

$$(\mathbf{Z}^T\mathbf{Z})^{-1} = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T$$

and apart from the factor σ^2, the variance of b_i is

$$\sum_{m=1}^{p} \frac{v_{im}^2}{\lambda_m}$$

so that the proportion of this variance which can be attributed to the eigenvalue λ_k is

$$P_{ki} = \frac{v_{ik}^2/\lambda_k}{\sum_{m=1}^{p} v_{im}^2/\lambda_m}.$$

The quantities P_{ki} are called the variance proportions.

Now, if the eigenvalue λ_k is small, this means that the corresponding eigenvector $\mathbf{v}_k$ describes a near linear dependence among the predictors (columns of $\mathbf{Z}$) since

$$\lambda_k = \mathbf{v}_k^T(\mathbf{Z}^T\mathbf{Z})\mathbf{v}_k = (\mathbf{Z}\mathbf{v}_k)^T(\mathbf{Z}\mathbf{v}_k)$$

and hence the vector formed by weighting the columns of $\mathbf{Z}$ by the elements of $\mathbf{v}_k$ is nearly the zero vector. The proportion of the variance of b_i that can be attributed to the linear dependence for a given eigenvalue λ_k is P_{ki}. If for a given small λ_k both P_{ki} and P_{kj} are large, this suggests that variables i and j are involved in a near linear dependence among the predictors. Our idea is to only have $\psi_{ij} \neq 0$ when both P_{ki} and P_{kj} are bigger than some cutoff value (0.25 say) for some eigenvalue λ_k. Once the ψ_{ij}^U are computed for pairs i, j for which $\psi_{ij} \neq 0$, we then follow the rescaling and truncation procedure of method B.

We also compare with a method based on that described by Denison et al. (1998), who used the reversible jump MCMC algorithm of Green (1995). We briefly describe our adaptation of their algorithm, since Denison et al. were concerned with nonparametric regression problems and adapting their method for general variable selection problems and a different prior specification requires a few changes.

The method of Denison, Mallick, and Smith (1998; hereafter DMS) proceeds by choosing between three different allowable move types at each step of the MCMC algorithm. Writing $\boldsymbol{\gamma}^{\text{cur}}$ for the current value of $\boldsymbol{\gamma}$, we generate a proposal $\boldsymbol{\gamma}^{\text{new}}$ by either (a) randomly choosing a component of $\boldsymbol{\gamma}^{\text{cur}}$ which is currently zero and making it one (birth step); (b) randomly choosing a component of $\boldsymbol{\gamma}^{\text{new}}$ which is currently one and making it zero (death step); or (c) simultaneously randomly choosing a component of $\boldsymbol{\gamma}^{\text{cur}}$ which is currently one and making it zero, and a component of $\boldsymbol{\gamma}^{\text{cur}}$ which is currently zero and making it one (flip step). For a model including k terms, we write b_k, d_k, and f_k for the probabilities of birth, death, and a flip, respectively. Denison et al. suggested choosing for $k = 1, 2, \ldots, p-1$

$$b_k = c \min\left\{1, \frac{p(k+1)}{p(k)}\right\},$$

and

$$d_k = c \min\left\{1, \frac{p(k)}{p(k+1)}\right\},$$

where c is a constant and $p(k)$ denotes the prior probability of a model including k terms (which in our case can be calculated from $p(\gamma)$). Of course, $f_k = 1 - b_k - d_k$. Also, we have $b_0 = 1$, $d_0 = f_0 = 0$ and $d_p = 1$, $b_p = f_p = 0$. The constant c must be chosen positive and so that $f_k \geq 0$ for all k: Denison et al. suggested $c = 0.4$ and we have chosen this value in our examples.

If a birth is proposed when there are k terms in the current model, then the acceptance probability is

$$\min\left\{\frac{p(\boldsymbol{\gamma}^{\text{new}}|\mathbf{y})d_{k+1}(n-k)}{p(\boldsymbol{\gamma}^{\text{cur}}|\mathbf{y})b_k(k+1)}\right\}.$$

For a death, the acceptance probability is

$$\min\left\{1, \frac{p(\boldsymbol{\gamma}^{\text{new}}|\mathbf{y})b_{k-1}k}{d_k(n-k+1)}\right\}$$

and for a flip the acceptance probability is simply

$$\min\left\{1, \frac{p(\boldsymbol{\gamma}^{\text{new}}|\mathbf{y})}{p(\boldsymbol{\gamma}^{\text{cur}}|\mathbf{y})}\right\}.$$

The modified DMS method is of interest for comparison with our method here because it allows a limited block move (through the “flip” step) which may offer improvements over one at a time sampling schemes in the case of multicollinearity.

6. EXAMPLES

Our applications involving simulated data are based on an example described by George and McCulloch (1997) and we have followed their approach to comparing sampling schemes by computing Monte Carlo standard errors of estimated marginal probabilities of inclusion for each of the predictors for the various sampling schemes. Write $\overline{\gamma}_i$, $i = 1, \ldots, p$ for the estimated marginal probabilities of inclusion. The Monte Carlo standard errors of these values are

$$\text{SE}(\overline{\gamma}_i) = \left[\frac{1}{k}\sum_{|h|<k}\left(1 - \frac{|h|}{k}\right)R_i(h)\right]^{1/2},$$

where $R_i(h)$ is the estimated autocovariance function of the sequence of iterates for γ_i assuming stationarity and k is the number of iterates: in practice, the autocovariances will be close to zero beyond some lag so that the sum can be truncated. In fact, truncation of the sum is necessary when the autocovariance function is estimated rather than known to ensure consistency of the normalized estimator: see, for instance, Besag and Green (1993). Because we run two chains for each method to check that the results obtained from different starting points are the same, the Monte Carlo standard errors reported in our tables are averages of the values from the two sequences.

It can be argued that the Monte Carlo standard errors should not be compared based on an equal number of iterations for all methods, but rather on the basis of equal processing time. In the following examples there was essentially no difference between time taken per iteration for the four different methods, since the computationally intensive part of each of the algorithms is the evaluation of the likelihood function, which occurs once per iteration for each of the methods. However, in problems with a large number of predictors, such as the example in Section 6.3, the time taken to obtain the initial estimate of the interaction parameters ψ_{ij} is substantial for method B. For methods C and D there are no substantial initializing computations.

FORTRAN code for implementing our methods can be obtained by e-mailing the first author.

6.1 A SIMULATED DATASET

Our first example was discussed by George and McCulloch (1997). They simulated a dataset with 15 predictor variables as follows. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_{15}, \mathbf{Z} \sim N_{180}(0, \mathbf{I})$, where $N_m(0, \mathbf{I})$ denotes the m-dimensional normal distribution with mean vector zero and covariance matrix $\mathbf{I}$. Then let $\mathbf{X}_i = \mathbf{Z}_i + 2\mathbf{Z}$, $i = 1, 3, 5, 8, 9, 10, 12, 13, 14, 15$ and set $\mathbf{X}_2 = \mathbf{X}_1 + 0.15\mathbf{Z}_2, \mathbf{X}_4 = \mathbf{X}_3 + 0.15\mathbf{Z}_4, \mathbf{X}_6 = \mathbf{X}_5 + 0.15\mathbf{Z}_6, \mathbf{X}_7 = \mathbf{X}_8 + \mathbf{X}_9 - \mathbf{X}_{10} + 0.15\mathbf{Z}_7$ and $\mathbf{X}_{11} = \mathbf{X}_{14} + \mathbf{X}_{15} - \mathbf{X}_{12} - \mathbf{X}_{13} + 0.15\mathbf{Z}_{11}$. George and McCulloch pointed out that this construction results in severe and complicated muticollinearity: there is a correlation of about 0.998 between $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$, $i = 1, 3, 5$ and strong linear dependencies among $(\mathbf{X}_7, \mathbf{X}_8, \mathbf{X}_9, \mathbf{X}_{10})$ and $(\mathbf{X}_{11}, \mathbf{X}_{12}, \mathbf{X}_{13}, \mathbf{X}_{14}, \mathbf{X}_{15})$. Let $\mathbf{X}$ be the design matrix with columns $\mathbf{X}_i$, $i = 1, \ldots, 15$. Let

$$\boldsymbol{\beta} = (1.5, 0, 1.5, 0, 1.5, 0, 1.5, -1.5, 0, 0, 1.5, 1.5, 1.5, 0, 0)^T,$$

and generate the responses $\mathbf{Y}$ as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N_{180}(0, 2.5^2\mathbf{I})$. This is the simulated data we have used to compare our sampling schemes.

For this simulated dataset, we ran 50,000 iterations of our sampling schemes from two starting points for the chain (the starting points were the model including all predictors and the model including none of them) and for the four methods (A, B, C, and DMS) described in the previous section. We discarded a burn-in period of 1,000 iterates for each sequence.

Table 1 shows Monte Carlo standard errors of estimated marginal probabilities of inclusion for the 15 predictors in the model for the four sampling schemes. From the table, it seems that methods B, C, and DMS all give an improvement over method A; recall that method A is provably better than the Gibbs sampler. Methods B and C are most promising.

Table 1. Monte Carlo Standard Errors for $\overline{\gamma}_i$ for George and McCulloch Simulated Dataset for Methods A, B, C, and DMS. Estimates are based on 50,000 iterations from two different starting points for each sampling scheme with 1,000 iterations burn in. The columns labeled "Relative" for methods B, C, and DMS give relative improvements of the Monte Carlo standard errors for these methods compared to that for method A.

	Method						
	A	*B*		*C*		*DMS*	
Predictor	*SE($\overline{\gamma}_i$)*	*SE($\overline{\gamma}_i$)*	*Relative*	*SE($\overline{\gamma}_i$)*	*Relative*	*SE($\overline{\gamma}_i$)*	*Relative*
X_1	0.0051	0.0031	1.65	0.0032	1.62	0.0050	1.01
X_2	0.0101	0.0077	1.31	0.0074	1.38	0.0089	1.13
X_3	0.0304	0.0075	4.05	0.0083	3.66	0.0217	1.40
X_4	0.0282	0.0069	4.06	0.0072	3.94	0.0204	1.39
X_5	0.0021	0.0017	1.24	0.0017	1.20	0.0027	0.76
X_6	0.0092	0.0088	1.05	0.0093	0.99	0.0095	0.97
X_7	0.0545	0.0078	6.94	0.0082	6.69	0.0418	1.31
X_8	0.0537	0.0080	6.67	0.0080	6.71	0.0411	1.31
X_9	0.0542	0.0070	7.74	0.0067	8.03	0.0412	1.31
X_{10}	0.0543	0.0072	7.60	0.0069	7.82	0.0413	1.31
X_{11}	0.0355	0.0080	4.44	0.0088	4.06	0.0295	1.20
X_{12}	0.0365	0.0077	4.72	0.0074	4.97	0.0287	1.27
X_{13}	0.0365	0.0075	4.83	0.0076	4.80	0.0297	1.23
X_{14}	0.0343	0.0046	7.47	0.0052	6.67	0.0266	1.29
X_{15}	0.0331	0.0046	7.29	0.0056	5.92	0.0263	1.26

Our method produces efficiency gains by encouraging "flips" between variables which are involved in a multicollinearity which are currently different, and these kind of flips seem to greatly improve mixing in this example.

6.2 Two Larger Simulated Datasets

We have also experimented with two variations on the example of George and McCulloch (1997) discussed in the previous subsection where the number of potential predictors is increased to 30. With 30 predictors, direct calculation of the normalizing constant in the posterior is difficult, and unlike the previous example MCMC methods really do become necessary for exploring the posterior distribution.

In the first variation of George and McCulloch's example we generate design matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ as in the previous example, but now for a sample size of 300. Then we let $\mathbf{X} = [\mathbf{X}_1\, \mathbf{X}_2]$ so that $\mathbf{X}$ is a 300 by 30 design matrix. Also, let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be 15 by 1 vectors both equal to the vector $\boldsymbol{\beta}$ used in the previous example, and let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T\, \boldsymbol{\beta}_2^T]^T$. Then we consider a dataset generated as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N_{300}(0, 2.5^2\mathbf{I})$. The efficiency of the sampling schemes is compared in the same way as for the previous example, but now 200,000 iterations of each sampler were used for each run. The results are shown for the first 15 predictor variables in Table 2. For brevity we have reported the results for only the first 15 predictor variables because of the symmetry in the way that the first and last 15 predictors and their coefficients are constructed. Again methods B and C seem superior to methods A and DMS.

Table 2. Monte Carlo Standard Errors for $\overline{\gamma}_i$ for Simulated Example with 30 Predictors for Methods A, B, C, and DMS. Estimates are based on 200,000 iterations from two different starting points for each sampling scheme with 1,000 iterations burn in. The columns labeled "Relative" for methods B, C, and DMS give relative improvements of the Monte Carlo standard errors for these methods compared to that for method A.

	Method						
	A	*B*		*C*		*DMS*	
Predictor	*SE*($\overline{\gamma}_i$)	*SE*($\overline{\gamma}_i$)	*Relative*	*SE*($\overline{\gamma}_i$)	*Relative*	*SE*($\overline{\gamma}_i$)	*Relative*
X_1	0.0251	0.0054	4.64	0.0039	6.51	0.0230	1.09
X_2	0.0248	0.0057	4.30	0.0042	5.96	0.0223	1.11
X_3	0.0207	0.0064	3.23	0.0060	3.41	0.0172	1.20
X_4	0.0209	0.0066	3.19	0.0057	3.70	0.0176	1.19
X_5	0.0197	0.0055	3.54	0.0057	3.48	0.0135	1.46
X_6	0.0203	0.0071	2.88	0.0060	3.38	0.0143	1.42
X_7	0.0561	0.0045	12.48	0.0045	12.48	0.0510	1.10
X_8	0.0561	0.0040	14.01	0.0046	12.05	0.0519	1.08
X_9	0.0559	0.0053	10.66	0.0049	11.30	0.0513	1.09
X_{10}	0.0551	0.0037	14.68	0.0055	10.01	0.0512	1.08
X_{11}	0.0811	0.0046	17.43	0.0049	16.71	0.0498	1.63
X_{12}	0.0812	0.0050	16.08	0.0058	14.00	0.0485	1.67
X_{13}	0.0813	0.0055	14.65	0.0055	14.78	0.0486	1.67
X_{14}	0.0804	0.0039	20.62	0.0042	19.14	0.0472	1.70
X_{15}	0.0817	0.0040	20.41	0.0042	19.67	0.0475	1.72

Table 3. Monte Carlo Standard Errors for $\overline{\gamma}_i$ for Simulated Example with 30 Predictors for Methods A, B, C, and DMS. Estimates are based on 200,000 iterations from two different starting points for each sampling scheme with 1,000 iterations burn in. The columns labelled "Relative" for methods B, C, and DMS give relative improvements of the Monte Carlo standard errors for these methods compared to that for method A.

	Method						
	A	*B*		*C*		*DMS*	
Predictor	*SE*($\overline{\gamma}_i$)	*SE*($\overline{\gamma}_i$)	*Relative*	*SE*($\overline{\gamma}_i$)	*Relative*	*SE*($\overline{\gamma}_i$)	*Relative*
X_1	0.0148	0.0047	3.12	0.0051	2.91	0.0112	1.33
X_2	0.0159	0.0051	3.12	0.0055	2.89	0.0117	1.36
X_3	0.0265	0.0071	3.73	0.0044	6.01	0.0173	1.53
X_4	0.0278	0.0083	3.35	0.0049	5.63	0.0180	1.55
X_5	0.0126	0.0030	4.20	0.0055	2.28	0.0087	1.45
X_6	0.0139	0.0033	4.15	0.0065	2.15	0.0094	1.47
X_7	0.0867	0.0059	14.82	0.0043	20.40	0.0556	1.56
X_8	0.0871	0.0058	15.02	0.0049	17.95	0.0563	1.55
X_9	0.0780	0.0064	12.19	0.0035	22.29	0.0504	1.55
X_{10}	0.0786	0.0059	13.44	0.0057	13.91	0.0516	1.52
X_{11}	0.0142	0.0037	3.79	0.0024	5.80	0.0162	0.88
X_{12}	0.0152	0.0031	4.90	0.0034	4.46	0.0162	0.94
X_{13}	0.0152	0.0027	5.53	0.0024	6.20	0.0163	0.93
X_{14}	0.0125	0.0014	9.26	0.0013	9.54	0.0149	0.83
X_{15}	0.0115	0.0013	9.20	0.0012	9.58	0.0151	0.76

Table 4. Predictors for U.S. Crime Dataset

Predictor	*Description*
M	percentage of males aged 14–24
So	indicator variable for a southern state
Ed	mean years of schooling
Po1	police expenditure in 1960
Po2	police expenditure in 1959
LF	labour force participation rate
M.F	number of males per 1000 females
Pop	state population
NW	number of nonwhites per 1,000 people
U1	unemployment rate of urban males 14–24
U2	unemployment rate of urban males 35–39
GDP	gross domestic product per head
Ineq	income inequality
Prob	probability of imprisonment
Time	average time served in state prisons

We also looked at another example identical to the one just described but where the last 15 components of $\boldsymbol{\beta}$ were set to zero (so that $\boldsymbol{\beta}$ has a fairly small number of nonzero components). Comparison between methods is shown in Table 3. Again we report results for the first 15 predictor variables: here the last 15 predictors have only small posterior probabilities of inclusion. Again methods B and C seem greatly superior. As it stands, in problems like this one where there are many useless predictors all the methods considered here will spend a lot of computational effort attempting to update predictors which have nearly zero probability of inclusion in the model. Kohn et al. (2001) proposed a sampling scheme which is designed to deal with this situation of many useless predictors where proposal values are sampled from the conditional prior. We note that we could modify our algorithm to use a similar Metropolis-Hastings proposal for clusters.

These examples involve 30 potential predictors: we believe that if the size of the problem was scaled up further then similar benefits could be obtained through the use of our algorithm, although this very much depends on the nature of the data.

6.3 U.S. Crime Rates

As a second example we consider a dataset on U.S. crime rates discussed by Ehrlich (1973). See also Raftery, Madigan, and Hoeting (1997). Interest in this example is in describing the relationship between the crime rate in 47 states of the U.S. and a set of predictors including measures describing sentencing regimes. The response is the rate of crimes in a particular category per head of population, and there are 15 predictors which are listed in Table 4. The predictors police expenditure in 1960 and police expenditure in 1959 are highly correlated, as are the predictors unemployment rate of urban males 14–24 and unemployment rate of urban males 35–39.

Table 5 shows Monte Carlo standard errors of $\overline{\gamma}_i$ for the 15 predictors. Methods B, C, and DMS all indicate an improvement over method A, although the gains are not as great as in the previous simulated examples, perhaps because the multicollinearities are not as severe.

Table 5. Monte Carlo Standard Errors for $\overline{\gamma}_i$ for U.S. Crime Example for methods A, B, C, and DMS. Estimates are based on 50,000 iterations from two different starting points for each sampling scheme with 1,000 iterations burn in. The columns labeled "Relative" for methods B, C, and DMS give relative improvements of the Monte Carlo standard errors for these methods compared to that for method A.

	Method						
	A	*B*		*C*		*DMS*	
Predictor	*SE($\overline{\gamma}_i$)*	*SE($\overline{\gamma}_i$)*	*Relative*	*SE($\overline{\gamma}_i$)*	*Relative*	*SE($\overline{\gamma}_i$)*	*Relative*
M	0.0120	0.0110	1.10	0.0118	1.03	0.0112	1.08
So	0.0074	0.0065	1.14	0.0070	1.06	0.0064	1.16
Ed	0.0120	0.0111	1.09	0.0142	0.85	0.0143	0.84
Po1	0.0197	0.0072	2.72	0.0079	2.49	0.0129	1.52
Po2	0.0222	0.0085	2.62	0.0092	2.43	0.0142	1.57
LF	0.0082	0.0065	1.26	0.0080	1.03	0.0060	1.38
M.F	0.0121	0.0093	1.30	0.0117	1.03	0.0109	1.12
Pop	0.0062	0.0072	0.85	0.0078	0.79	0.0058	1.06
NW	0.0065	0.0072	0.90	0.0072	0.90	0.0060	1.07
U1	0.0074	0.0058	1.28	0.0078	0.95	0.0055	1.35
U2	0.0082	0.0072	1.13	0.0087	0.94	0.0089	0.91
GDP	0.0074	0.0075	0.99	0.0089	0.84	0.0081	1.91
Ineq	0.0072	0.0060	1.18	0.0089	0.80	0.0087	0.82
Prob	0.0095	0.0117	0.82	0.0106	0.90	0.0102	0.94
Time	0.0079	0.0074	1.07	0.0076	1.05	0.0065	1.22

6.4 Statistical Correction of a Numerical Weather Prediction Model

Our third example concerns a regression model for statistical correction of a deterministic numerical weather prediction model. The responses consist of 369 observations of daily maximum temperatures at Sydney airport throughout August, September, and October 1993–1996. There are 62 predictors in our dataset which are averages of 24-hour and 36-hour forecasts of 62 meteorological fields obtained from a numerical weather prediction model. Many of the numerical weather prediction model predictors are closely related to each other and so this dataset is one that involves a very large number of predictors and severe multicollinearity. For more background on the data see Nott, Dunsmuir, Kohn, and Woodcock (2001). A similar procedure to the previous examples was followed for comparing methods, but this time 200,000 iterations were obtained for each chain.

Table 6 shows Monte Carlo standard errors of $\overline{\gamma}_i$ for a number of the predictors X_i. These predictors were the ones which had estimated marginal probability of inclusion of between 0.15 and 0.85 based on the results of all sampling schemes. It appears that methods C and DMS again outperform method A, although once more the advantage is not as decisive as in the simulated examples. The DMS method appears best of the four approaches here. The length of time taken to obtain the initial estimates of the ψ_{ij} in method B is a disadvantage of this method: the time taken to compute the interaction parameters is more than the time taken to compute the MCMC iterates! In comparison, for method C, computing the ψ_{ij} takes just a few seconds.

Table 6. Monte Carlo Standard Errors for $\overline{\gamma}_i$ for Numerical Weather Prediction Model Data for Methods A, B, C, and DMS. Estimates are based on 200,000 iterations for each method from two different starting methods with 1,000 iterations burn in for each sequence. The columns labeled "Relative" for methods B, C, and DMS give relative improvements of the Monte Carlo standard errors for these methods compared to that for method A.

	Method						
	A	*B*		*C*		*DMS*	
Predictor	*SE*($\overline{\gamma}_i$)	*SE*($\overline{\gamma}_i$)	*Relative*	*SE*($\overline{\gamma}_i$)	*Relative*	*SE*($\overline{\gamma}_i$)	*Relative*
X_6	0.0452	0.0508	0.89	0.0473	0.96	0.0229	1.98
X_8	0.0608	0.0401	1.51	0.0321	1.90	0.0271	2.24
X_{26}	0.0384	0.0466	0.82	0.0450	0.85	0.0155	2.48
X_{31}	0.0192	0.0213	0.90	0.0092	2.07	0.0117	1.64
X_{50}	0.0542	0.0221	2.46	0.0210	2.58	0.0219	2.48
X_{56}	0.0251	0.0159	1.58	0.0172	1.46	0.0174	1.45

7. DISCUSSION AND CONCLUSIONS

This article describes a sampling scheme for Bayesian variable selection which is based on the Swendsen-Wang algorithm for the Ising model and which can perform better than currently used sampling schemes in problems where there are multicollinearities amongst the predictors.

We mention briefly one potentially interesting extension of the present work. One common application of Bayesian variable selection methods with a large number of potential predictors is to nonparametric regression using linear combinations of basis functions. Kohn et al. (2001) developed sampling schemes more efficient than traditional sampling schemes for this problem. When describing the mean response function in terms of a linear combination of a large number of basis functions and where most of the basis functions are not needed we effectively have a variable selection problem with many useless predictors. Kohn et al. (2001) suggested Metropolis-Hastings schemes for which the Metropolis-Hastings acceptance ratio will be fast to compute when updating components of γ corresponding to useless predictors.

There is the potential to employ similar ideas in our sampling scheme in applications to nonparametric regression. Furthermore, in some nonparametric regression problems there would be a natural way of choosing which of the interaction parameters ψ_{ij} are nonzero in our algorithm. Consider the bivariate regression model

$$y_i = f(x_i, z_i) + \epsilon_i,$$

where y_i is the ith response, x_i and z_i are values of two predictor variables, and the ϵ_i are zero mean normal constant variance errors. There are many possible choices for a basis that can flexibly approximate the function f. For instance, one choice is a thin plate spline basis: writing $\mathbf{t} = (x, z)$ and $\mathbf{t}_1, \ldots, \mathbf{t}_r$ for a collection of knot points, the thin plate basis is

$$\{1, x, z, \|\mathbf{t} - \mathbf{t}_1\|^2 \log(\|\mathbf{t} - \mathbf{t}_1\|), \ldots \|\mathbf{t} - \mathbf{t}_r\|^2 \log(\|\mathbf{t} - \mathbf{t}_r\|)\},$$

where $\|\cdot\|$ is the Euclidean norm (see, for instance, Green and Silverman 1994, chap. 7). We can choose the knots as the observed predictor values, or we could do a cluster analysis of the predictor vectors to get a more parsimonious set of knots. In expanding the mean function in terms of this basis, we have indicator variables γ_i associated with each of the knot points and we can think very naturally of the γ_i's as a spatial field with spatial indices given by the knots. We could allow a nonzero ψ_{ij} in our algorithm only for pairs γ_i, γ_j corresponding to knot points which are close to each other. Basis functions corresponding to nearby knot points are likely to be similar, leading to multicollinearity in the design matrix. Hence our sampling scheme could be more efficient than currently used sampling schemes for this problem.

[Received October 2001. Revised October 2002.]

REFERENCES

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation" (with discussion), *Journal of the Royal Statistical Society*, Series B, 16, 395–407.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society*, Series B, 60, 333–350.

Ehrlich, I. (1973), "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation," *Journal of Political Economy*, 81, 521–565.

George, E. I., and McCulloch, R. E. (1997), "Approaches to Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

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

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.

Higdon, D. M. (1998), "Auxiliary Variable Methods for Markov Chain Monte Carlo With Applications," *Journal of the American Statistical Association*, 93, 585–595.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial" (with discussion), *Statistical Science*, 14, 382–417; corrected version available at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Kohn, R., Smith, M., and Chan, D. (2001), "Nonparametric Regression Using Linear Combinations of Basis Functions," *Statistics and Computing*, 11, 313–322.

Madigan, D., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

Myers, R. H. (1990), *Classical and Modern Regression with Applications* (2nd ed.), Belmont, CA: Duxbury.

Nott, D. J., Dunsmuir, W. T. M., Kohn, R., and Woodcock, F. (2001), "Statistical Correction of a Deterministic Numerical Weather Prediction Model," *Journal of the American Statistical Association*, 96, 794–804.

Peskun, P. H. (1973), "Optimum Monte Carlo Sampling Using Markov Chains," *Biometrika*, 60, 607–612.

Raftery, A. E. (1996), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," *Biometrika*, 83, 251–266.

Raftery, A. E., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–344.

Swendsen, R. H., and Wang, J. S. (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physical Review Letters*, 58, 86–88.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997), "Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke," *Journal of the Royal Statistical Society*, Series C, 46, 433–448.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions," in *Bayesian Inference and Decision Techniques—Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, Amsterdam: North Holland, pp. 233–243.