Bayesian variable selection and the Swendsen-Wang algorithm

David J. Nott[1] and Peter J. Green[2]

1. Department of Statistics
University of New South Wales
Sydney 2052, Australia
e-mail: djn@maths.unsw.edu.au

2. Department of Mathematics
University of Bristol
Bristol BS8 1TW, UK.
e-mail: P.J.Green@bristol.ac.uk

Abstract

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. Traditional sampling algorithms for Bayesian variable selection may perform poorly when there are severe multicollinearities amongst the predictors. In this paper we describe a new sampling method based on an analogy with the Swendsen-Wang algorithm for the Ising model, and which can give substantial improvements over traditional sampling schemes in the presence of multicollinearity. In linear regression with a given set of potential predictors we can index different possible models by a binary parameter vector which 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 approximate the posterior distribution by an Ising model and then apply a modified Swendsen-Wang algorithm for sampling 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.

Keywords: Bayesian variable selection, Markov chain Monte Carlo, Swendsen-Wang algorithm.

1 Introduction

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

$$y = X\beta + \epsilon$$

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

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 *et al.* (1998), Kohn *et al.* (2001) and the references therein for a discussion of different approaches and recent developments.

Traditional sampling schemes for exploring the posterior distribution may be slow mixing when there are severe multicollinearities amongst the predictors. In this paper, we describe an algorithm which offers improvements over traditional schemes in this situation, and which is based on an analogy with the Swendsen-Wang algorithm for the Ising model. We formulate our hierarchical prior for β in terms of a vector of binary variables in which the components indicate whether a predictor is included in the model or not. We think of this binary parameter vector as a spatial process, approximate its posterior distribution by an Ising model, and then employ an algorithm similar to the Swendsen-Wang algorithm for sampling from the posterior distribution in which 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).

The structure of the paper is as follows. In Section 2 we specify the model and the priors. In Section 3 we review the Swendsen-Wang algorithm, and we extend the algorithm to the problem of Bayesian variable selection in Section 4. Section 5 describes our method for approximating the posterior distribution on the model space via an Ising model. Section 6 describes performance of the method for real and simulated data, and Section 7 gives some discussion and conclusions.

2 Bayesian variable selection

As in the introduction, write y for a vector of n responses, X for an $n \times p$ design matrix and β for a vector of regression coefficients. Let $\gamma = (\gamma_1, ..., \gamma_p)^T$ be a binary vector, and write $q_\gamma = \sum_i \gamma_i$ for the number of nonzero elements of γ. Let X_γ be the $n \times q_\gamma$ design matrix obtained by removing those columns i from X for which $\gamma_i = 0$. Similarly let β_γ be the subvector of β obtained by removing components β_i of β for which $\gamma_i = 0$.

We assume that

$$y|\gamma, X_\gamma, \beta_\gamma, \sigma^2 \sim N(X_\gamma\beta_\gamma, \sigma^2 I).$$

For Bayesian inference on the model parameters we use a hierarchical prior. The prior for β_γ given γ and σ^2 is normal,

$$p(\beta_\gamma|\gamma, \sigma^2) \quad \sim \quad N(0, n\sigma^2(X_\gamma^T X_\gamma)^{-1}). \tag{1}$$

The prior on σ^2 is $p(\sigma^2) \propto \sigma^{-2}$, and for our prior on γ we use $p(\gamma) = 2^{-p}$, so that all models have equal prior probability. For alternative prior specifications on γ that encourage model parsimony see Denison *et al.* (1998) and Kohn *et al.* (2001).

We are interested in the posterior distribution on γ with β_γ and σ^2 integrated out,

$$p(\gamma|y) \quad \propto \quad p(\gamma)p(y|\gamma). \tag{2}$$

Here it can be shown that

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

For p relatively small, we can compute the posterior $p(\gamma|y)$ exactly, obtaining the normalizing constant in (2) by summing over all possible values of γ. 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 γ the usual Markov chain Monte Carlo methods for exploring the posterior, which update one component of γ at a time, can be slow mixing. High posterior correlations can occur, for instance, in the situation where there is multicollinearity. Updating components of γ 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.

We describe an alternative auxiliary variable technique for efficiently exploring the posterior distribution when there are high posterior correlations. The method is based on an analogy with the Swendsen-Wang algorithm (Swendsen and Wang, 1987) which is used to sample from the Ising model and its generalizations in statistical physics. A review of the Swendsen-Wang algorithm and discussion of some extensions to general problems in Bayesian inference is given by Higdon (1998).

3 Swendsen-Wang algorithm

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

$$p(\eta) \quad \propto \quad \exp\left(\sum_i \alpha_i(\eta_i) + \sum_{i<j} \beta_{ij} I(\eta_i = \eta_j)\right) \tag{3}$$

where $\beta_{ij} \geq 0, i < j$, and $I(A)$ is the indicator function which is one when A occurs and zero otherwise. Ratios of the expression (3) are easy to compute for η 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 be very slow mixing when there is strong dependence between components of η. There are critical values for the interaction parameters β_{ij} around which a small change can result in a large change in the global behaviour of $p(\eta)$, and for values of β_{ij} near these critical values single site udpating schemes may not work well.

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 η. We let

$$u = \{u_{ij} : 1 \leq i < j \leq p\}$$

be a set of auxiliary variables and set up a joint distribution $p(u, \eta)$ on u and η in which the marginal distribution for η is given by (3). This joint distribution can be constructed so that $p(u|\eta)$ and $p(\eta|u)$ are easy to sample from.

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

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

Then we have

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

and of course $p(\eta|u) \propto p(u,\eta)$.

If $\beta_{ij} > 0$, then the condition $u_{ij} < \exp(\beta_{ij} I(\eta_i = \eta_j))$ is satisfied if $u_{ij} < 1$, or even if $u_{ij} \in [1, \exp(\beta_{ij})]$ if $\eta_i = \eta_j$. So when $u_{ij} \in [1, \exp(\beta_{ij})]$, η_i and η_j are constrained to be equal, 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 $p(\eta|u)$ that components of η are conditionally independent. In effect, conditioning on the auxiliary variables removes the interactions between the components of η. Hence both $p(u|\eta)$ and $p(\eta|u)$ are easy to sample from.

4 Analogy with Bayesian variable selection

By thinking about the posterior distribution (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.

Suppose that some initial approximation to $p(\gamma|y)$ can be obtained of the form (3) where now we allow the interaction parameters β_{ij} to be less than zero. We discuss how to choose the interaction parameters later. Now let $u = \{u_{ij} : 1 \le i < j \le p\}$ be a collection of auxiliary variables, and define a joint distribution $p(u, \gamma|y)$ by specifying the distribution of γ given y to be the posterior distribution $p(\gamma|y)$ of Section 2 and the conditional distribution $p(u|\gamma, y)$ to be

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

That is, the u_{ij}'s are conditionally independent given γ and uniform as in the Swendsen-Wang algorithm.

Then

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

Although $p(\gamma|y)$ does not in general have the form (3), it is hoped that as in the Swendsen-Wang algorithm the denominator in the above expression will serve to reduce interactions among components of γ conditional on u. This is the key idea in our method.

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

The auxiliary variables u_{ij} define clusters among the components of γ in much the same way as in the Swendsen-Wang algorithm. Components of γ 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 γ. Let $C = C(u)$ be one cluster defined by the set of auxiliary variables u, and let $\overline{C}$ denote the set of variables not in C. Let $\gamma(C)$ be the subset of γ corresponding to the variables in C, and $\gamma(\overline{C})$ denote the remaining components of γ. Given the constraints, note that there are only two possible values for the vector $\gamma(C)$: 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 $\gamma(C)$ by a Metropolis-Hastings step. Write γ^{new} for a proposed value of γ in which $\gamma^{new}(C)$ is generated from the proposal distribution $q(\gamma(C)|\gamma, u, y)$ and $\gamma^{new}(\overline{C}) = \gamma(\overline{C})$. The Metropolis-Hastings acceptance probability is

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

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

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

where

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

This is inexpensive to compute provided that β_{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 $\gamma(C)|\gamma(\overline{C}), u, y$, which makes the Metropolis-Hastings acceptance probability equal to one. If we set $\beta_{ij} = 0$ 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

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

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

$$\min\left(1, \frac{1-q}{q}\right). \tag{4}$$

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 only consider 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 β_{ij} in $p(u|\gamma, y)$. For computational reasons it is advisable to keep the number of nonzero β_{ij} as small as possible (so that the Metropolis-Hastings acceptance probabilities are inexpensive to compute). Also, the nonzero β_{ij} should not be too large in magnitude, as pointed out by Higdon (1998) with respect to some problems in image analysis, for the following reason. If many of the $|\beta_{ij}|$ 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 γ_i in a large cluster will have a value fixed by the likelihood so that a proposal to "flip" is unlikely to be accepted. Hence if we have many large $|\beta_{ij}|$, encouraging large clusters, it is difficult for the sampling scheme to move from the current configuration for the γ's.

Our proposed method for choosing the interaction parameters β_{ij} is based on the following idea. Let $\eta = (\eta_1, ..., \eta_p)^T$ be a binary spatial process with joint distribution of the form (3). Let $\eta^* = (\eta_1^*, ..., \eta_p^*)^T$ be some fixed configuration for the sites. Then the following formula holds, regardless of the value chosen for η^*:

$$\beta_{ij} \;=\; 0.5 \times \left(\sum_{\eta_i=\eta_j\,,\,\eta_k=\eta_k^*,k\neq i,j} \log p(\eta) - \sum_{\eta_i\neq\eta_j\,,\,\eta_k=\eta_k^*,k\neq i,j} \log p(\eta) \right). \tag{5}$$

Note that this formula can still be applied even when $\log p(\eta)$ is only known up to an additive constant.

We can use this formula or some variant of it to obtain an approximation to the posterior distribution $p(\gamma|y)$ of the required form for partial decoupling. In particular, for some binary vector $\gamma^* = (\gamma_1^*, ..., \gamma_p^*)^T$ we compute

$$\beta_{ij}^U \;=\; 0.5 \times \left(\sum_{\gamma_i=\gamma_j\,,\,\gamma_k=\gamma_k^*,k\neq i,j} \log p(\gamma|y) - \sum_{\gamma_i\neq\gamma_j\,,\,\gamma_k=\gamma_k^*,k\neq i,j} \log p(\gamma|y) \right). \tag{6}$$

and then obtain β_{ij} by rescaling β_{ij}^U, $\beta_{ij} = c\beta_{ij}^U$ where c is a scaling factor chosen so that the resulting β_{ij} are not too large. The rescaling is necessary for the reason pointed out earlier: if large clusters grow, then at least one of the γ_i in the cluster is likely to have a value fixed by the likelihood, and any proposal to "flip" is unlikely to be accepted. Modifications of the Swendsen-Wang algorithm that involve scaling down of the interaction parameters have been previously considered (see Hidon, 1998, for example). In the Swendsen-Wang algorithm for the model (3), growth of clusters is based only on the values of the interaction parameters β_{ij} without regard to the values of the terms $\alpha_i(\cdot)$ or which sites η_i might be fixed by the model. This is detrimental to the performance of the Swensden-Wang algorithm, and some scaling down of the interaction parameter values when defining the auxiliary variables may be beneficial (see Higdon for further discussion and some examples).

After rescaling the parameters obtained in (6) it may also be useful to incorporate a thresholding operation so that any β_{ij} small in magnitude are set to zero. We note that the value obtained for β_{ij}^U in (6) will in general depend on the value γ^*, since $p(\gamma|y)$ is not generally of the form (3). One way of avoiding this dependence is to average over all possible models, taking

$$\beta_{ij}^U \;=\; 2^{-(p-1)} \left(\sum_{\gamma:\gamma_i=\gamma_j} \log p(\gamma|y) - \sum_{\gamma:\gamma_i\neq\gamma_j} \log p(\gamma|y) \right). \tag{7}$$

Incidentally, this formula also gives a way of determining the parameters β_{ij} in a generalization of (3) in which higher order interaction parameters are included.

However, applying (7) is infeasible when there are large numbers of predictors due to the number of terms in the summations.

We investigate in the empirical studies of the next section four basic rules for determining the interaction parameters. Our first rule is to simply set $\beta_{ij} = 0$ 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 for the following reason. In Section 3 we mentioned that the Gibbs sampler results from setting $\beta_{ij} = 0$ in our method when clusters are updated by the cluster full conditionals. If we implement our method with the antithetic proposal and $\beta_{ij} = 0$, then the antithetic modification results in an algorithm which is provably better, by Peskun (1973) Theorem 2.1.1.

Our second rule for choosing interaction parameters, hereafter method B, is based on determining β_{ij}^U according to (7) and then rescaling these values so that they lie between negative one and one. We obtain final values for β_{ij} after rescaling the β_{ij}^U by setting to zero any values which are less than 0.1 in absolute value. That is, let

$$\beta_{ij}^S = \frac{\beta_{ij}^U}{\max_{i,j} |\beta_{ij}^U|}$$

and then set

$$\beta_{ij} = \beta_{ij}^S I(|\beta_{ij}^S| \geq 0.1).$$

We have experimented with scaling the β_{ij}^U to lie in different ranges. Scaling to the range $[-a, a]$ for any value of a between 0.25 and 2.5 seems to give an improvement over $\beta_{ij} = 0$ in a range of problems. As mentioned earlier, Higdon (1998) considers a similar rescaling idea in defining auxiliary variables for a modification of the Swendsen-Wang algorithm, and offers suggestions about how the rescaling can be done in some problems in image analysis. Experimentation and fine tuning with the scaling obviously helps to optimize efficiency in our application, but we use a scaling to the interval $[-1, 1]$ as an automatic rule, and this generally seems to work well.

For our third rule, method C, we allow γ^* in (6) to be a vector of ones: that is, we consider the full model. Intuitively, setting all components of γ^* 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. See Section 7 for further dicussion. The values of β_{ij}^U obtained are transformed as for method B: they are rescaled to lie between negative one and one, and the rescaled values which are less than 0.1 in absolute value are set to zero to obtain the final β_{ij} values.

For our fourth rule, method D, we also allow γ^* in (6) to be a vector of ones, but we now allow only a much smaller number of the β_{ij} to be nonzero. Computing

β_{ij}^U for all pairs i and j when there is a large number of predictors in method C can be very time consuming. So if we can reduce the number of pairs i, j for which we must compute the interaction parameter β_{ij} then this can improve computational efficiency.

Our method for choosing which pairs i, j have an interaction parameter $\beta_{ij} \neq 0$ involves the use of a standard multicollinearity diagnostic, the variance proportions. For further background see, for instance, Myers (1991). We try to identify severe linear dependencies among columns of the design matrix using the variance proportions, and only allow $\beta_{ij} \neq 0$ for those columns i, j involved in a severe linear dependence.

We now describe the method more precisely. Let Z be the $n \times p$ matrix obtained from the design matrix X by centring and scaling each of the predictors (so that each column of 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 centred. Write

$$Z^T Z = V D V^T$$

for the eigenvalue decomposition of $Z^T Z$, where $V = [v_1, ..., v_p]$ is the orthogonal matrix with columns given by the eigenvectors $v_1, ..., v_p$ of $Z^T Z$ and D is the diagonal matrix of eigenvalues with diagonal entries $\lambda_1, ..., \lambda_p$. The least squares estimate b of $\beta = (\beta_1, ..., \beta_p)^T$ in the model

$$y = Z\beta + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$ has covariance matrix

$$\sigma^2 (Z^T Z)^{-1}.$$

We can write

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

where D^{-1} is the inverse of D, the diagonal matrix with ith diagonal entry $1/\lambda_i$. Apart from the factor σ^2, the variance of b_i is

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

and 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 v_k describes a near linear dependence among the predictors (columns of Z) since

$$\lambda_k = v_k^T (Z^T Z) v_k = (Z v_k)^T (Z v_k)$$

and hence the vector formed by weighting the columns of Z by the elements of 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 small 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 which impacts detrimentally on estimation of coefficients for these predictors. Our idea is to only have $\beta_{ij} \neq 0$ when both P_{ki} and P_{kj} are bigger than some cutoff value (0.5 say) for some eigenvalue λ_k. Once the β_{ij}^U are computed for pairs i, j for which $\beta_{ij} \neq 0$, we rescale them to lie between -1 and 1 and set scaled values less than 0.1 in absolute value to zero.

We have experimented with many other techniques for choosing the interaction parameters, but the methods described above were most successful.

6 Examples

We describe the performance of our simulation algorithms in three examples.

6.1 A simulated example

Our first example was discussed by George and McCulloch (1997). They simulated a data set with 15 predictor variables as follows. Let $Z_1, ..., Z_{15}, Z \sim N_{180}(0, I)$, where $N_m(0, I)$ denotes the m-dimensional normal distribution with mean vector zero and covariance matrix I. Then let $X_i = Z_i + 2Z$, $i = 1, 3, 5, 9, 10, 12, 13, 14, 15$ and set $X_2 = X_1 + 0.15Z_2$, $X_4 = X_3 + 0.15Z_4$, $X_6 = X_5 + 0.15Z_6$, $X_7 = X_8 + X_9 - X_{10} + 0.15Z_7$ and $X_{11} = X_{14} + X_{15} - X_{12} - X_{13} + 0.15Z_{11}$. George and McCulloch point out that this construction results in severe and complicated muticollinearity: there is a correlation of about 0.998 between X_i and X_{i+1}, $i = 1, 3, 5$ and strong linear dependencies among (X_7, X_8, X_9, X_{10}) and $(X_{11}, X_{12}, X_{13}, X_{14}, X_{15})$. Let X be the design matrix with columns X_i, $i = 1, ..., 15$. Let

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

and generate the responses Y as

$$Y = X\beta + \epsilon$$

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

For this simulated data set, 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 D) of choosing the interaction parameters described in the previous section. We discarded a burn in period of 1000 iterates for each sequence.

Following George and McCulloch's approach for comparing different sampling schemes, we computed Monte Carlo standard errors of the estimated marginal posterior probabilities of inclusion for each of the variables. Write $\overline{\gamma}_i$, $i = 1, ..., 15$ for the estimated marginal probabilities of inclusion. The Monte Carlo standard errors of these values are

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

where $R_i(h)$ is the estimated autocovariance function of the sequence of iterates for γ_i 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 to ensure consistency of the estimator: see, for instance, Besag and Green (1993). Since 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.

Table 1 shows these Monte Carlo standard errors for 13 of the 15 predictors in the model for the three sampling schemes. No standard errors are reported for X_1 and X_5 since these predictors have an estimated marginal probability of inclusion close to one based on the iterates from all sampling schemes, so that the corresponding Monte Carlo standard errors are nearly zero for all the methods. From the table, it seems that methods B, C and D all give an improvement over method A, and recall that method A is provably better than the Gibbs sampler. Methods C and D are most promising.

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 this example there was essentially no difference between time taken per iteration for the three different 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 β_{ij} is substantial for methods B and C.

6.2 US crime rates

As a second example we consider a data set on US crime rates discussed by Ehrlich (1973). See also Raftery (1995). Interest in this example is in describing the relationship between the crime rate in 47 states of the US 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 2. 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 4 shows Monte Carlo standard errors of $\overline{\gamma}_i$ for the 15 predictors. Methods B, C and D all indicate an improvement over method A, although the gains are not as great as in the previous simulated example, perhaps because the multicollinearities are not as severe. Methods C and D appear to be better than Method B.

6.3 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 data set 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 data set is one that involves a very large number of predictors and severe multicollinearity. For more background on the data see Nott *et al.* (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 3 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. Method B cannot feasibly be implemented in a problem this large, so results for this method are not shown. It appears that methods C and D again outperform method A, although once more the advantage is not as decisive as in the simulated example. The length of time taken to obtain the initial estimates of the β_{ij} in method C 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 here! In comparison, for method D, computing the β_{ij} takes just a few seconds.

7 Discussion and Conclusions

In this paper we have described a sampling scheme for Bayesian variable selection which is based on the Swendsen-Wang algorithm for the Ising model and which performs better than traditional sampling schemes when there are multicollinearities amongst the predictors.

We mention briefly a few modifications and extensions of the methods we have presented. Firstly, we have implemented our method for a different prior specification to the one considered so far. In particular, we have considered using the prior of Kohn *et al.* (2001) for γ,

$$p(\gamma|\pi) = \pi^{q_\gamma}(1-\pi)^{p-q_\gamma}$$

where π is a hyperparameter with a beta prior,

$$p(\pi) = \frac{1}{B(a,b)}\pi^{a-1}(1-\pi)^{b-1}\, 0 \leq \pi \leq 1$$

where $B(\cdot,\cdot)$ is the beta function. The prior on γ (integrating out π) is

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

We have found that for this prior specification we need to modify our method for choosing the interaction parameters β_{ij} in our algorithm, as we now explain.

Write $\gamma_{\neq i,j}$ for the set of all elements of γ excluding γ_i and γ_j. Clearly we can rewrite (6) in the form

$$\beta_{ij}^U \quad = \quad 0.5 \times \left(\sum_{\gamma_i=\gamma_j} \log p(\gamma_i, \gamma_j | y, \gamma_{\neq i,j} = \gamma^*_{\neq i,j}) \right.$$

$$- \sum_{\gamma_i \neq \gamma_j} \log p(\gamma_i, \gamma_j | y, \gamma_{\neq i,j} = \gamma^*_{\neq i,j}) \Bigg) . \tag{8}$$

If the posterior distribution is of the form (3), then the value obtained in the expression above for $\beta^U_{i,j}$ will not depend on the value of $\gamma^*_{\neq i,j}$. However, in general, the posterior is not of the form (3) and we want to choose the $\beta^U_{i,j}$ so that for our approximation of the form (3) to the posterior the conditional distribution for γ_i, γ_j given $\gamma_{\neq i,j} = \gamma^*_{\neq i,j}$ matches the corresponding true posterior conditional distribution well for all possible values of $\gamma^*_{\neq i,j}$. We believe that it is most important for the approximation to work well for values of $\gamma^*_{\neq i,j}$ which are typical of samples from the posterior distribution of $\gamma_{\neq i,j}$. In the prior specification we have considered in this paper, it seems as though choosing $\gamma^*_{\neq i,j}$ as a vector of ones helps us to capture fairly well the conditional relationship between γ_i and γ_j in a model typical of a sample from the posterior distribution. However, when we used the prior of Kohn *et al.* (2001) this did not seem to be the case, for the following reason.

Note that we can write

$$p(\gamma_i, \gamma_j | y, \gamma_{\neq i,j}) \propto p(\gamma_i, \gamma_j | \gamma_{\neq i,j}) p(y|\gamma)$$

so that we can rewrite (8) as

$$\begin{aligned} \beta^U_{ij} &= 0.5 \times \left(\sum_{\gamma_i = \gamma_j} \log p(\gamma_i, \gamma_j | \gamma_{\neq i,j} = \gamma^*_{\neq i,j}) - \sum_{\gamma_i \neq \gamma_j} \log p(\gamma_i, \gamma_j | \gamma_{\neq i,j} = \gamma^*_{\neq i,j}) \right) \\ &\quad + 0.5 \times \left(\sum_{\gamma_i = \gamma_j, \gamma_{\neq i,j} = \gamma^*_{\neq i,j}} \log p(y|\gamma) - \sum_{\gamma_i \neq \gamma_j, \gamma_{\neq i,j} = \gamma^*_{\neq i,j}} \log p(y|\gamma) \right) \end{aligned} \tag{9}$$

So we can separate out prior and likelihood contributions to β^U_{ij}. If we set $\gamma^*_{\neq i,j}$ 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 (9) and base calculation of β^U_{ij} only on the likelihood $p(y|\gamma)$ with $\gamma^*_{\neq i,j}$ a vector of ones. Note that we are in effect recommending that we use the same method for determining the interaction parameters as we used for our original prior specification, since when all models are equally likely the prior contribution in the above expression disappears.

We also describe 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) develop 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) suggest 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 β_{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 $t = (x, z)$ and $t_1, ..., t_r$ for a collection of knot points, the thin plate basis is

$$\{1, x, z, \|t - t_1\|^2 \log(\|t - t_1\|), ...\|t - t_r\|^2 \log(\|t - t_r\|)\}$$

where $\| \cdot \|$ is the Euclidean norm (see, for instance, Green and Silverman, 1994, Chapter 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 β_{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 traditional sampling schemes for this problem.

References

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.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998), "Automatic Bayesian curve fitting", *Journal of the Royal Statistical Society*, Ser. 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. and Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.

Higdon, D.M. (1998), "Auxiliary variable methods for Markov chain Monte Carlo with applications", *Journal of the American Statistical Association*, 93, 585-595.

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

Myers, R.H. (1990), *Classical and Modern Regression with Applications (Second Edition)*, Duxbury, Belmont, California.

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 - Applications and Case Studies*, conditionally accepted.

Peskun, P.H. (1973), "Optimum Monte Carlo sampling using Markov chains", *Biometrika*, 60, 607-612.

Raftery, A.E. (1995), "Bayesian model selection in social research," In: *Sociological Methodology 1995*, ed. P.V. Marsden, pp. 111-196. Oxford: Blackwells.

	Method						
Predictor	A	B		C		D	
	SE($\overline{\gamma}_i$)	SE($\overline{\gamma}_i$)	Relative	SE($\overline{\gamma}_i$)	Relative	SE($\overline{\gamma}_i$)	Relative
X_2	0.0080	0.0086	0.9244	0.0068	1.1691	0.0063	1.2619
X_3	0.0294	0.0097	3.0625	0.0080	3.6750	0.0090	3.2667
X_4	0.0288	0.0106	2.7170	0.0084	3.4286	0.0077	3.7403
X_6	0.0077	0.0100	0.7750	0.0092	0.8424	0.0062	1.2500
X_7	0.0799	0.0445	1.7966	0.0075	10.6600	0.0090	8.8833
X_8	0.0796	0.0457	1.7418	0.0060	13.2667	0.0077	10.3377
X_9	0.0797	0.0447	1.7841	0.0064	12.4609	0.0073	10.9247
X_{10}	0.0783	0.0456	1.7171	0.0063	12.4286	0.0080	9.7875
X_{11}	0.0707	0.0261	2.7088	0.0065	10.8769	0.0079	8.9494
X_{12}	0.0721	0.0285	2.5316	0.0080	9.0187	0.0080	9.0187
X_{13}	0.0713	0.0278	2.5647	0.0075	9.5067	0.0063	11.3175
X_{14}	0.0685	0.0260	2.6346	0.0049	13.9796	0.0052	13.1731
X_{15}	0.0693	0.0259	2.6766	0.0045	15.4111	0.0047	14.7553

Table 1: Monte Carlo standard errors for $\overline{\gamma}_i$ for simulated example for methods A, B, C and D. Estimates are based on 50,000 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 D give relative improvements of the Monte Carlo standard errors for these methods compared to that for method A.

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 1000 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

Table 2: Predictors for US crime data set

	Method						
Predictor	A	B		C		D	
	SE($\overline{\gamma}_i$)	SE($\overline{\gamma}_i$)	Relative	SE($\overline{\gamma}_i$)	Relative	SE($\overline{\gamma}_i$)	Relative
M	0.0113	0.0095	1.1895	0.0069	1.6259	0.0075	1.4967
So	0.0124	0.0108	1.1535	0.0073	1.6871	0.0070	1.7714
Ed	0.0102	0.0089	1.1397	0.0105	0.9761	0.0092	1.1027
Po1	0.0147	0.0142	1.0389	0.0118	1.2511	0.0106	1.3868
Po2	0.0155	0.0090	1.7127	0.0077	2.0000	0.0078	1.9745
LF	0.0177	0.0122	1.4508	0.0076	2.3289	0.0092	1.9239
M.F	0.0100	0.0089	1.1236	0.0081	1.2422	0.0095	1.0471
Pop	0.0115	0.0124	0.9274	0.0109	1.0550	0.0101	1.1443
NW	0.0079	0.0092	0.8587	0.0089	0.8827	0.0076	1.0327
U1	0.0097	0.0066	1.4809	0.0071	1.3759	0.0073	1.3288
U2	0.0099	0.0093	1.0645	0.0088	1.1250	0.0077	1.2857
GDP	0.0099	0.0092	1.0820	0.0058	1.7069	0.0089	1.1061
Ineq	0.0086	0.0079	1.0886	0.0100	0.8643	0.0095	0.9053
Prob	0.0105	0.0077	1.3725	0.0078	1.3642	0.0087	1.2139
Time	0.0111	0.0105	1.0571	0.0063	1.7619	0.0084	1.3136

Table 3: Monte Carlo standard errors for $\overline{\gamma}_i$ for US crime example for methods A, B, C and D. Estimates are based on 50,000 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 D give relative improvements of the Monte Carlo standard errors for these methods compared to that for method A.

Predictor	Method				
	A	C		D	
	SE($\overline{\gamma}_i$)	SE($\overline{\gamma}_i$)	Relative	SE($\overline{\gamma}_i$)	Relative
X_6	0.0237	0.0234	1.0107	0.0183	1.2951
X_{26}	0.0194	0.0205	0.9463	0.0160	1.2125
X_{31}	0.0153	0.0125	1.2240	0.0178	0.8596
X_{49}	0.0158	0.0138	1.1449	0.0101	1.5644
X_{56}	0.0152	0.0186	0.8172	0.0195	0.7815
X_{57}	0.0128	0.0090	1.4144	0.0118	1.0894

Table 4: Monte Carlo standard errors for $\overline{\gamma}_i$ for numerical weather prediction model data for methods A, C and D. 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 labelled “Relative” for methods C and D give relative improvements of the Monte Carlo standard errors for these methods compared to that for method A.