Adaptive Monte Carlo for Binary Regression with Many Regressors

D. Lamnisos* J. E. Griffin† and M. F. J. Steel*

Abstract

This article describes a method for efficient posterior simulation for Bayesian variable selection in probit regression models with many regressors but few observations. A proposal on model space is described which contains a tuneable parameter. An adaptive approach to choosing this tuning parameter is described which allows automatic, efficient computation in these models. The methods is applied to the analysis of gene expression data.

1 INTRODUCTION

There are many problems that amount to the selection of a few variables, from a much larger set, with the aim of discriminating between two classes. For example, measurements of gene expression may be taken in a microarray experiment with the goal of finding a small subset of the genes that allow discrimination between two conditions such as disease or non-diseased, or two types of cancer. The standard approach compares the gene expression levels for the two groups on a gene-by-gene basis (see e.g. Dudoit et al, 2002). In this paper the problem is seen as variable selection in a probit regression model in a Bayesian framework (see e.g. Lee et al (2003), Sha et al (2003, 2004) and Yeung et al (2005)). This approach takes into account correlations between the variables and provides a measure of uncertainty in the choice of variables which also extends to any predictions. A Markov chain Monte Carlo (MCMC) method is used for posterior inference and our goal is to automatically produce a Markov chain with good mixing properties which gives accurate answers with the smallest possible run length. Alternative methods of estimation have been proposed. Variational methods for binary regression models with Gaussian priors have been discussed by e.g. Jaakola and Jordan (2000). Variable selection using sparsity-inducing priors have been discussed by e.g. Qi and Jaakola (2007).

Markov chain Monte Carlo (MCMC) has become the main tool for simulating from a target distribution, $\pi(x)$, which in our case will be a posterior distribution. One of the simplest such methods is the Random Walk Metropolis-Hastings (RWM) sampler, which generates a Markov chain whose marginal distribution is $\pi(x)$. Suppose that $x_n$ is the current value of the chain, a potential new value of the chain, $x'$, is proposed as
\[ x' = x_n + \epsilon_n \] where \( \epsilon_n \) is a random perturbation drawn from some distribution (often taken to be normal) with mean zero and standard deviation \( \sigma \). The potential value is accepted as the next value of the chain \( x_{n+1} = x' \) or rejected \( x_{n+1} = x_n \) with a probability given by the Metropolis-Hastings acceptance ratio. This method is popular because careful tuning of \( \sigma \) leads to the optimal RWM sampler and it can be combined with Gibbs sampling to produce simulation schemes for a wide-range of models. For many target distributions, the optimal RWM sampler occurs when the average acceptance probability \( \bar{\tau} \) is 0.234 (see e.g. Roberts and Rosenthal, 2001 and Sherlock and Roberts, 2009). Tuning of the algorithm is simple since \( \bar{\tau} \) will typically fall as the standard deviation \( \sigma \) is increased. The average acceptance rate \( \bar{\tau} \) can be estimated from pilot runs with a fixed value of \( \sigma \) and the standard deviation can be adjusted so that the estimated \( \bar{\tau} \) is close to 0.234.

Recently, there has been interest in adaptive Monte Carlo methods where the distribution of the proposal, \( x' \), is adjusted during the MCMC run. These methods are difficult to implement in general since the Markov property is violated and standard theory for convergence of the chain to the target distribution does not apply. However, convergence to the target distribution can be verified for particular forms of adjustment. The first adaptive algorithm that could be shown to converge to the target distribution was introduced by Haario et al (2001) who used methods from Stochastic Approximation. This important idea and other methods are reviewed by Andrieu and Thoms (2008).

This paper is concerned with extending these methods to problems in variable selection with many regressors. The posterior distribution will be complicated and vast (there will be \( 2^p \) potential models if there are \( p \) potential regressors). Adaptive methods are important because MCMC methods often mix slowly (and so proposals that encourage good mixing are important) and the running of many pilot runs is unsatisfactory due to the large number of iterations needed to give good estimates of posterior summaries. Such methods have been previously applied to variable selection problems by Nott and Kohn (2005). They allow the probability that a particular variable is proposed to be included in or removed from the model to adapt over the chain. This is rather different to the method developed here where the number of variables, rather than which variables, is adapted over the chain. The method uses a form of proposal for variable selection described by Lamnisos et al (2009) which can be tuned in a similar way to an RWM sampler.

The paper is organised as follows: Section 2 describes the Bayesian approach to variable selection and some MCMC algorithms for posterior exploration, Section 3 describes a tuneable proposal for variable selection problems, Section 4 describes adaptive versions of the algorithms, Section 5 includes some numerical examples that demonstrate the utility of the approach and Section 6 includes some concluding comments.

## 2 BAYESIAN VARIABLE SELECTION

Let \( y_1, y_2, \ldots, y_n \) be observations taking values 0 and 1 where the \( i \)-th observation is associated with a set of \( p \) regressors \( x_{i1}, x_{i2}, \ldots, x_{ip} \). In this paper, it is assumed that \( p \) is much larger than \( n \). For example, \( y_i \) could be the disease status of the \( i \)-th patient and \( x_{i1}, x_{i2}, \ldots, x_{ip} \) is a vector of gene expression measurements or proteomic measurements (see e.g. Lee et al (2003), Sha et al (2003, 2004) and Yeung et al (2005)). It is assumed that only a subset of the regressors are needed to predict \( y_i \) and we use the indicator
variables $\gamma_1, \gamma_2, \ldots, \gamma_p$ to represent whether the $i$-th regressor is included in the model ($\gamma_i = 1$ if a regressor is included and $\gamma_i = 0$ otherwise). Let $p_\gamma = \sum_{i=1}^{p} \gamma_i$ be the size of model $\gamma$. It is assumed that observations are generated by the probit model:

$$p(y_i = 1) = \Phi(\alpha + x_i^T \beta_\gamma),$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution, $x_i^\gamma$ is a vector containing only the elements of $x_i$ for which $\gamma_i = 1$ and $\beta_\gamma$ is a $p_\gamma$-dimensional vector of regression coefficients. We also define $X_\gamma$ to be the submatrix of $X$ which only includes those columns for which $\gamma_i = 1$. Within a Bayesian analysis, both $\beta_\gamma$ and $\gamma$ are given prior distributions. The regression coefficients are assumed to be $\beta_\gamma | \gamma \sim N(\mathbf{0}, V_\gamma)$ where $V_\gamma$ might follow the so-called $g$-prior form $V_\gamma = c(X_\gamma^T X_\gamma)^{-1}$ or the independent form $V_\gamma = cI$ where $I$ is the identity matrix and $c$ is a positive scalar. The prior assumes that $p(\gamma_i = 1) = w$ which implies that the number of included variables is binomially distributed with mean $nw$ and variance $nw(1 - w)$. The hyperparameter $w$ can be interpreted as the prior probability of including a variable in the model.

The posterior distribution of $\gamma$ and $\theta_\gamma = (\alpha, \beta_\gamma)$ is not available analytically and so computational methods are needed. Markov chain Monte Carlo methods are a popular class of algorithms to fit the model. Several methods have been proposed in the literature. Holmes and Held (2006) describe one algorithm, which uses data augmentation with a vector $z$ based on a representation described in Albert and Chib (1993):

1. Generate $z = (z_1, z_2, \ldots, z_n)'$ where $z_i \sim N(\alpha + x_i^\gamma \beta_\gamma, 1)$ where $z_i > 0$ if $y_i = 1$ or $z_i < 0$ if $y_i = 0$.
2. Select model $\gamma'$ with probability $q(\gamma'|\gamma)$.
3. Jump to the model $\gamma'$ with probability

$$\alpha(\gamma, \gamma') = \min \left\{1, \frac{\pi(\gamma') q(\gamma'|\gamma) \pi(z|\gamma')}{\pi(\gamma) q(\gamma'|\gamma) \pi(z|\gamma)} \right\}.$$

4. If the jump is accepted, draw a sample $\theta_{\gamma'} \sim N((X_{\gamma'}^T X_{\gamma'} + V_{\gamma'}^{-1})^{-1} X_{\gamma'}^T z, (X_{\gamma'}^T X_{\gamma'} + V_{\gamma'}^{-1})^{-1})$.

Alternatively, one can update $\gamma$ and $\theta_\gamma$ jointly. Green (2003) describes an Automatic Generic Sampler which is extended by Lamnisos et al (2009).

1. Draw a sample $\theta_\gamma \sim N((X_\gamma^T X_\gamma + V_\gamma^{-1})^{-1} X_\gamma^T z, (X_\gamma^T X_\gamma + V_\gamma^{-1})^{-1})$.
2. Select model $\gamma'$ with probability $q(\gamma'|\gamma)$.
3. Propose $\theta_{\gamma'}$ in the following way. Let $\mu_\gamma$ and $\Sigma_\gamma$ be an approximation of the mean and variance of the posterior distribution of $\theta_\gamma$ and let $B_\gamma$ be the Cholesky decomposition of $\Sigma_\gamma$ and $v = B_\gamma^{-1}(\theta_\gamma - \mu_\gamma)$. Then we propose $\theta_{\gamma'} = \mu_\gamma + B_\gamma v'$ where

$$v' = \begin{cases} (v_1, \ldots, v_{p_{\gamma'}})^T & \text{if } p_{\gamma'} < p_\gamma \\ (v^T, \epsilon^T)^T & \text{if } p_{\gamma'} = p_\gamma \\ (v^T, \epsilon^T)^T & \text{if } p_{\gamma'} > p_\gamma \end{cases}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_{p_{\gamma'}-p_\gamma})^T$ has i.i.d. $N(0, 1)$ elements.
4. Jump to the model $\gamma'$ and parameter $\theta_{\gamma'}$ with probability $\alpha(\gamma, \gamma', \theta_{\gamma}, \theta_{\gamma'}) = \min\left\{ 1, \frac{\pi(\gamma', \theta_{\gamma'})q(\gamma' \mid \gamma')p(y \mid \theta_{\gamma'})|B_{\gamma'}|}{\pi(\gamma, \theta)q(\gamma \mid \gamma)p(y \mid \theta_{\gamma})|B_{\gamma}|} \times K \right\}$

$$K = \begin{cases} (2\pi)^{-\frac{1}{2}(p_{\gamma}-p_{\gamma'})} \exp \left\{ -\frac{1}{2} (\epsilon')^T (\epsilon') \right\} & \text{if } p_{\gamma'} < p_{\gamma} \\ 1 & \text{if } p_{\gamma'} = p_{\gamma} \\ (2\pi)^{-\frac{1}{2}(p_{\gamma}-p_{\gamma'})} \exp \left\{ \frac{1}{2} \epsilon^T \epsilon \right\} & \text{if } p_{\gamma'} > p_{\gamma}, \end{cases}$$

and $\epsilon'$ is the obvious counterpart of $\epsilon$.

5. Generate $z_1, z_2, \ldots, z_n$ where $z_i \sim N(\alpha + x_i^T \beta, 1)$ where $z_i > 0$ if $y_i = 1$ or $z_i < 0$ if $y_i = 0$.

There are several methods of finding $\mu_{\gamma}$ and $\Sigma_{\gamma}$. Two are considered in this paper: the Laplace approximation and the Iterated Weighted Least Squares (IWLS) approximation for one iteration (see Lannios et al. (2009) for more details).

Alternative automatic methods for moving between models are described by Brooks et al. (2003) which are applied in this context by Lannios et al. (2009). In the case of proposals that increase the model size, the coefficient vector is completed with $u_{\gamma'}(v) = \mu + \sigma v$ where $v$ has a standard normal distribution. The Conditional Maximization method chooses $\mu$ to maximize the posterior distribution $\pi((\theta_{\gamma}, u_{\gamma}) \mid y)$ with respect to $u_{\gamma}$. The variance, $\sigma^2$, is chosen to ensure that the acceptance probability of the move with $u_{\gamma}(v) = \mu$ is 1 and is given by

$$\sigma^2 = \left( \frac{\pi(y \mid \theta_{\gamma}, \gamma) \pi(\gamma)}{\pi(y \mid \theta_{\gamma}, \mu, \gamma') \pi(\gamma')} \frac{1}{\exp \left\{ \frac{u^T(u - \mu)}{2 \sigma^2} \right\}} \right)^{-1/2}.$$

Here

$$\alpha(\gamma, \gamma', \theta_{\gamma}, \theta_{\gamma'}) = \min\left\{ 1, \frac{\pi(y \mid \theta_{\gamma'}, \gamma') \pi(\gamma')}{\pi(y \mid \theta_{\gamma}, \gamma) \pi(\gamma)}\right\} B$$

where

$$B = \exp \left\{ \frac{1}{2} \left[ v^T v - (\mu + \sigma v)^T (\mu + \sigma v) / c \right] \right\} \left( \frac{\sigma^2}{c} \right)^{-\frac{p_{\gamma'} - p_{\gamma}}{2}}.$$

The pseudocode representation of the Conditional Maximization method is as follows:

If at iteration $t$ the current state is $(\theta_{\gamma}^{(t)}, \gamma)$, then

1. Draw a sample $\theta_{\gamma} \sim N((X_{\gamma}^T X_{\gamma} + V_{\gamma}^{-1})^{-1} X_{\gamma}^T z, (X_{\gamma}^T X_{\gamma} + V_{\gamma}^{-1})^{-1})$.

2. Select model $\gamma'$ with probability $q(\gamma' \mid \gamma)$.

3. Determine the location $\mu$ and the scale $\sigma$ of the proposal random variable $u_{\gamma}$ as described above.

4. Generate $u_{\gamma} \sim N_{p_{\gamma'} - p_{\gamma}}(\mu, \sigma^2 I_{p_{\gamma'} - p_{\gamma}})$. 

4
5. Set $\theta_{\gamma'} = (\theta_1^{(t)} , u_\gamma)$.

6. Jump to the model $\gamma'$ and set $\theta_2^{(t+1)} = \theta_{\gamma'}$ with probability given by (2). Otherwise set $\theta_3^{(t+1)} = \theta_2^{(t)}$.

7. Generate $z_1, z_2, \ldots, z_n$ where $z_i \sim \mathcal{N}(\alpha + x_i^T \beta, 1)$ where $z_i > 0$ if $y_i = 1$ or $z_i < 0$ if $y_i = 0$.

### 3 A TUNEABLE PROPOSAL ON MODEL SPACE

Lamnisos et al (2009) propose a new general model proposal $q_\zeta(\gamma' | \gamma)$ which draws a new model in the following way:

1. A value $N(t)$ is generated from a Binomial distribution with $N - 1$ trials and success probability $\zeta$.

2. One of three possible moves: Add, Delete and Swap is chosen uniformly at random. If Add is selected then $N(t) + 1$ regressors are chosen to be added to those included in $\gamma$ to form $\gamma'$, if Delete is selected then $N(t) + 1$ regressors are chosen to be removed from the model and if Swap is selected then $N(t) + 1$ regressors are swapped without changing the model size (provided $p_\gamma \geq N(t) + 1$; if not, the Add step is chosen).

This model proposal combines local moves with more global ones by changing simultaneously a block of variables. Two parameters determine this proposal: $N$ is the maximum number of variables that can be changed from the current model $\gamma$ and $\zeta$ determines the degree of “localness” since the mean number of variables changed is $1 - \zeta + N\zeta$. The value of $N$ will usually be fixed and the parameter $\zeta$ chosen to control the mixing of the chain. The application of this proposal to microarray data by Lamnisos et al (2009) suggests that the optimum effective sample size is obtained when the average acceptance rate fall in the range 0.25 to 0.40. This is true for a wide-range of sampling schemes. Rather like RWM samplers, this optimal choice of acceptance rate can be achieved by carefully tuning the parameter $\zeta$ of the model proposal using a series of pilot runs. In each pilot run, the sampler is run for a chosen value of $\zeta$ and the average acceptance rate calculated. If the acceptance rate is too high then $\zeta$ is increased in the next run and if the acceptance rate is too low then $\zeta$ is decreased in the next run. However, this tuning process is typically a computationally expensive task since trial and error is required.

As an alternative solution, we consider adaptive MCMC algorithms which can automatically handle this parameter tuning. The problem is similar to adaptation in RWM samplers since there is a tuneable proposal and an optimal acceptance rate to be achieved. We will extend the Adaptive Random Walk Metropolis (ARWM) algorithm proposed by Atchadé & Rosenthal (2005) to the variable selection problem. The ARWM algorithm automatically finds the optimal scale parameter in the RWM algorithm that results in the optimal acceptance rate $\bar{\tau} = 0.234$. Therefore, our adaptive MCMC algorithms automatically find $\zeta$ such that the resulting acceptance rate falls in the range 0.25 to 0.4. A comprehensive survey of recent advances in adaptive MCMC methodology and their applications can be found in Rosenthal (2008).
4 THE ADAPTIVE ALGORITHM

The ARWM algorithm of Atchadé & Rosenthal (2005) sequentially adapts the scale parameter $\sigma$ of the RWM with a $d$-variate normal proposal density centred on the current value and with variance $\sigma^2 I$ and for which the stationary distribution is the positive continuous density $\pi(x)$. The entire past of the stochastic process is used to adapt the scale parameter $\sigma$. The resulting sequence of scale parameters $\{\sigma(t) : t \in \mathbb{N}\}$ converges to an optimal value that leads to the optimal acceptance rate $\bar{\tau} = 0.234$. Atchadé & Rosenthal (2005) fix values $\varepsilon_1$ and $A_1$ such that $0 < \varepsilon_1 < A_1$ and defined the set $\Delta = \{\sigma : \varepsilon_1 \leq \sigma \leq A_1\}$. They also assume that $\Delta$ contains a unique value $\sigma_{opt}$ which results in the optimal acceptance rate $\bar{\tau}$.

Then, they define the following function of $\sigma$

$$\rho(\sigma) = \begin{cases} 
\varepsilon_1 & \text{if } \sigma < \varepsilon_1 \\
\sigma & \text{if } \sigma \in \Delta \\
A_1 & \text{if } \sigma > A_1 
\end{cases}$$

(3)

The aim of this function is to contain the adaptive algorithm inside $\Delta$. Finally, they define a discount factor $s(t)$, which is a positive sequence of real numbers such that $s(t) = O(t^{-\lambda})$ for some constant $1/2 < \lambda < 1$. This assumption ensures the ergodicity of the ARWM algorithm. The simple choice $s(t) = a t^{-1}$ for some $a > 0$ will meet this condition. The pseudocode of the ARWM algorithm proceeds as follows:

If at iteration $t$ the current state is $x(t) \in \mathbb{R}^d$ and the scale parameter of the proposal density is $\sigma(t) \in \Delta$, then

1. Generate $y \sim N(x(t), (\sigma(t))^2 I)$.

2. Set $x(t+1) = y$ with probability

$$\alpha(x(t), y) = \min\left\{1, \frac{\pi(y)}{\pi(x(t))}\right\},$$

otherwise take $x(t+1) = x(t)$.

3. Compute

$$\sigma(t+1) = \rho(\sigma(t) + s(t)(\alpha(x(t), y) - \bar{\tau})).$$

(4)

The acceptance rate is monitored by (4). The algorithm decreases the scale parameter $\sigma(t+1)$ when the acceptance rate is small and increases $\sigma(t+1)$ when the acceptance rate is high. Atchadé & Rosenthal (2005) showed under certain assumptions that the generated stochastic process is ergodic with stationary distribution the positive continuous density $\pi(x)$.

Turning to the MCMC algorithms of Section 2 with model proposal $q(\gamma' | \gamma)$, the parameter $\zeta$ behaves like the scale parameter $\sigma$ of the RWM because values of $\zeta$ close to 0 yield more local moves and high acceptance rate and values of $\zeta$ close to 1 more global moves and small acceptance rate. Moreover, our applications to some gene expression datasets suggest an optimal acceptance rate $\bar{\tau}$ between 0.25 and 0.4. Therefore, we adopt ideas of the ARWM to develop adaptive version of each transdimensional MCMC sampler.
described in Section 2. In our case $\zeta \in [0, 1]$, thus values of $\epsilon_1$ and $A_1$ are chosen close to 0 and 1, respectively. In our applications of the adaptive algorithms, $\epsilon_1$ and $A_1$ are set to 0.01 and 0.99, respectively. The parameter $\zeta$ can be made adaptive by updating it at the $t$-th iteration in the following way, analogous to the RWM,

$$\zeta^{(t+1)} = \rho(\zeta^{(t)} + s^{(t)}(\alpha^{(t)} - \bar{\tau}))$$  (5)

where $\alpha^{(t)}$ is the acceptance probability at the $t$-th iteration of the chain.

All the algorithms of Section 2 can be made adaptive by updating $\zeta$ at each iteration using the recursion in (5). The pseudocode representation of the adaptive Holmes and Held algorithm adjusts the model proposal step and adds an extra step (step 4 below) in the corresponding non-adaptive algorithm. This pseudocode representation is as follows:

If at iteration $t$ the current state is $(\theta^{(t)}_\gamma, \gamma)$ and $\zeta^{(t)} \in [0, 1]$, then

1. Generate $z_1, z_2, \ldots, z_n$ where $z_i \sim \mathcal{N}(\alpha + x_i^T \beta, 1)$ where $z_i > 0$ if $y_i = 1$ or $z_i < 0$ if $y_i = 0$.
2. Select model $\gamma'$ with probability $q^{(t)}(\gamma'|\gamma)$.
3. Jump to the model $\gamma'$ with probability

$$\alpha(\gamma, \gamma') = \min \left\{ 1, \frac{\pi(\gamma')q(\gamma'|\gamma')\pi(z|\gamma')}{\pi(\gamma)q(\gamma'|\gamma)\pi(z|\gamma)} \right\}.$$  

4. Compute

$$\zeta^{(t+1)} = \rho(\zeta^{(t)} + s^{(t)}(\alpha(\gamma, \gamma') - \bar{\tau})).$$

5. If $\gamma'$ is accepted, draw a sample $\theta^{(t)}_{\gamma'} \sim \mathcal{N}(X^T_{\gamma'}X_{\gamma'}^{-1}X^T_{\gamma'}z, (X^T_{\gamma'}X_{\gamma'}^{-1}+V_{\gamma'}^{-1})^{-1})$.

We think it is a safe conjecture that the results regarding ergodicity with the correct stationary distribution of the ARWM algorithm carry over to our model selection setting. In Section 5 we present empirical evidence which suggests that these results hold.

5 SIMULATION RESULTS

The performance of the adaptive MCMC algorithms is evaluated using two microarray data sets. These are the Arthritis data (Sha et al., 2003) and the Colon Tumour data (Alon et al., 1999). Adaptive versions of all algorithms in Section 2 were tested and are denoted as follows:

1. ADH-H : Adaptive Holmes and Held algorithm
2. ADAG-LA : Adaptive automatic generic sampler with Laplace approximation
3. ADAG-IWLS : Adaptive automatic generic sampler with Iterated Weighted Least Squares approximation
Non-adaptive versions of the algorithms are indicated by dropping the first two letters “AD”. All the adaptive MCMC samplers start with initial parameter value $\zeta_0 = 0.5$ and use the positive sequence of real numbers $\{s^{(t)} = \zeta_0/t : t \in \mathbb{N}\}$. The number of iterations was 2,000,000, the burn-in period 100,000 and the thinning 10 resulting in an MCMC sample size $T$ of 190,000. Finally, we specify the value 0.3 as an optimal acceptance rate $\bar{\tau}$ because the optimum effective sample size of the MCMC algorithms that explore the model and parameter space of our problem is obtained when acceptance rates are between 0.25 and 0.4. Adopting $\bar{\tau} = 0.234$ instead makes very little difference to our results.

Table 1: The effective sample size ESS, the CPU time in seconds of the adaptive and non-adaptive algorithms with relative efficiencies of the non-adaptive algorithm over the adaptive algorithm for the Arthritis and Colon Tumour datasets

### Arthritis

<table>
<thead>
<tr>
<th>Method</th>
<th>ESS</th>
<th>CPU</th>
<th>R.E</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H ($\zeta = 0$)</td>
<td>35144</td>
<td>9795</td>
<td>0.99</td>
</tr>
<tr>
<td>H-H ($\zeta = 0.25$)</td>
<td>35489</td>
<td>9880</td>
<td>1.00</td>
</tr>
<tr>
<td>ADH-H</td>
<td>34902</td>
<td>9668</td>
<td></td>
</tr>
<tr>
<td>AG-LA ($\zeta = 0$)</td>
<td>58001</td>
<td>40135</td>
<td>0.75</td>
</tr>
<tr>
<td>AG-LA ($\zeta = 0.5$)</td>
<td>80701</td>
<td>40346</td>
<td>1.04</td>
</tr>
<tr>
<td>ADAG-LA</td>
<td>80853</td>
<td>41966</td>
<td></td>
</tr>
<tr>
<td>AG-IWLS ($\zeta = 0$)</td>
<td>50942</td>
<td>9781</td>
<td>0.74</td>
</tr>
<tr>
<td>AG-IWLS ($\zeta = 0.5$)</td>
<td>68233</td>
<td>9870</td>
<td>0.98</td>
</tr>
<tr>
<td>ADAG-IWLS</td>
<td>69126</td>
<td>9822</td>
<td></td>
</tr>
<tr>
<td>C-M ($\zeta = 0$)</td>
<td>55641</td>
<td>49173</td>
<td>0.95</td>
</tr>
<tr>
<td>C-M ($\zeta = 0.5$)</td>
<td>66801</td>
<td>49993</td>
<td>1.12</td>
</tr>
<tr>
<td>ADC-M</td>
<td>64243</td>
<td>53944</td>
<td></td>
</tr>
</tbody>
</table>

### Colon Tumour

<table>
<thead>
<tr>
<th>Method</th>
<th>ESS</th>
<th>CPU</th>
<th>R.E</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H ($\zeta = 0$)</td>
<td>32752</td>
<td>13968</td>
<td>0.99</td>
</tr>
<tr>
<td>ADH-H</td>
<td>32228</td>
<td>13587</td>
<td></td>
</tr>
<tr>
<td>AG-LA ($\zeta = 0$)</td>
<td>47354</td>
<td>41355</td>
<td>0.97</td>
</tr>
<tr>
<td>AG-LA ($\zeta = 0.25$)</td>
<td>53653</td>
<td>42345</td>
<td>1.08</td>
</tr>
<tr>
<td>ADAG-LA</td>
<td>54777</td>
<td>45716</td>
<td></td>
</tr>
<tr>
<td>AG-IWLS ($\zeta = 0$)</td>
<td>45494</td>
<td>14165</td>
<td>0.88</td>
</tr>
<tr>
<td>AG-IWLS ($\zeta = 0.25$)</td>
<td>51971</td>
<td>14246</td>
<td>1.00</td>
</tr>
<tr>
<td>ADAG-IWLS</td>
<td>51231</td>
<td>14066</td>
<td></td>
</tr>
<tr>
<td>C-M ($\zeta = 0$)</td>
<td>39596</td>
<td>48159</td>
<td>0.94</td>
</tr>
<tr>
<td>C-M ($\zeta = 0.25$)</td>
<td>42370</td>
<td>48313</td>
<td>1.00</td>
</tr>
<tr>
<td>ADC-M</td>
<td>41861</td>
<td>47789</td>
<td></td>
</tr>
</tbody>
</table>
The efficiency of an MCMC sampler can be measured using the Effective Sample Size (ESS) which is $T/(1 + 2 \sum_{j=1}^{\infty} \rho_j)$ for an MCMC run of length $T$ with lag $j$ autocorrelation $\rho_j$ (see e.g., Liu, 2001). The interpretation is that the MCMC sampler has the same accuracy of estimates as a Monte Carlo sampler (where all the draws are independent) run for ESS iterations. In this paper, the ESS is estimated using the Initial Positive Sequence Estimator (Geyer, 1992). The algorithms have different running times and so we define the efficiency ratio for a sampler to be

$$ER(\text{Sampler}) = \frac{\text{ESS(Sampler)}}{\text{CPU(Sampler)}},$$

which standardizes the effective sample size by CPU run time and so penalizes computationally inefficient algorithms. We are interested in the performance of each adaptive algorithms to the non-adaptive algorithm with $\zeta = 0$ (which is the standard MCMC proposal for these types of models and represents a baseline) and with the optimal value of $\zeta$ among five candidates ($\zeta = 0, 0.25, 0.5, 0.75, 0.95$) which results in the highest ER. The relative efficiency of the non-adaptive over the adaptive algorithm is defined by

$$R.E = \frac{ER(\text{Non-Adaptive})}{ER(\text{Adaptive})}.$$

Table 1 presents result of the adaptive algorithms and various non-adaptive algorithms with fixed values of $\zeta$ for the Arthritis and Colon Tumour datasets. The relative efficiency for all sampling methods against the standard proposal ($\zeta = 0$) is always less than 1 indicating that the adaptive method is superior. The increase in performance depends on the particular method and the form of the posterior. However, the effect can be large in some cases. For example, the standard method only obtains 75% of the efficiency of the adaptive method with the Arthritis data and AG-LA and AG-IWLS algorithms. It is also clear that the Automatic Generic methods gain the most benefit. In fact, the effective sample size of the adaptive algorithms are very similar to the optimal non-adaptive algorithms in terms of mixing. Furthermore, the increase in CPU time of the adaptive algorithms is small. This leads to relative efficiencies quite close to 1 and therefore the adaptive algorithms achieve essentially the same efficiency as the optimal non-adaptive algorithms. Crucially, however, the adaptive algorithms avoid the pilot runs needed to tune the model proposal parameter $\zeta$.

Figure 1 and Figure 2 show the trace plots of both the model proposal parameter $\zeta$ (left panels) and the empirical acceptance rate (right panels) of the adaptive algorithms for the Arthritis and Colon Tumour datasets, respectively. The parameter $\zeta$ of each adaptive algorithm converges to a value close to the optimal one obtained by manual tuning. Furthermore, the empirical acceptance rates converge to values quite close to the target acceptance rate 0.3. These results illustrate that the adaptive MCMC algorithms automatically find model proposal parameters $\zeta$ that give asymptotically the optimal acceptance rate $\bar{\tau} = 0.3$.

Figure 3 displays the scatter-plots of the log estimated posterior gene inclusion probabilities of the adaptive and optimal non-adaptive algorithms for the Arthritis and Colon Tumour datasets. The log posterior gene inclusion probabilities are very similar indicating empirically that the stationary distribution of the stochastic process generated by the adaptive MCMC algorithms is the target joint posterior distribution $\pi(\theta, \gamma | y)$. 

9
6 DISCUSSION

This paper describes an adaptive Monte Carlo algorithm for posterior simulation for variable selection in probit regression models with many regressors. The algorithm leads to Markov chains with good mixing properties and without the need for pilot runs. In fact, the effective sample sizes for the adaptive algorithms are almost identical to those for the algorithms run at an optimized value of the proposal parameter $\zeta$ (found using trial-and-error). The methods are useful when there are a large number of variables that could potentially be included in the model, which leads to high average acceptance rates for standard algorithms. If the number of regressors is not large, then acceptance rates will not be high and an average acceptance rate of 0.3 may not be achievable. In this case, $\zeta$ should be close to zero and the value of $\zeta^{(t)}$ will converge to a value close to zero showing the robustness of the algorithm. Therefore, we suggest the use of adaptive MCMC algorithms to explore efficiently the model space of Bayesian variable selection in probit regression with many covariates. High acceptance rates for standard MCMC algorithms
in variable selection are also observed with other models such as linear regression when there are many regressors. The application of these adaptive methods to other models is an area of future research.

Acknowledgements

Demetris Lamnisos would like to acknowledge support from the Centre for Research in Statistical Methodology (CRiSM) at the University of Warwick.

References


Figure 3: Scatter-plots of the log estimated posterior gene inclusion probabilities of the adaptive and optimal non-adaptive algorithms for the Arthritis and Colon Tumour datasets

and Normal Colon Tissues Probe by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences of the USA* 96: 6745-6750.


