Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations

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

March 7, 2008

Abstract

One flexible technique for model search in probit regression is Markov chain Monte Carlo methodology that simultaneously explores the model and parameter space. The reversible jump sampler is designed to achieve this simultaneous exploration. Standard samplers, such as those based on MC^3, often have low model acceptance probabilities when there are many more regressors than observations. Simple changes to the form of the proposal leads to much higher acceptance rates. However, high acceptance rates are often associated with poor mixing of chains. This suggests defining a more general model proposal that allows us to propose models "further" from our current model. We design such a proposal which can be tuned to achieve a suitable acceptance rate for good mixing (rather like the tuning of a random walk proposal in fixed dimension problems). The effectiveness of this proposal is linked to the form of the marginalisation scheme when updating the model and we propose a new efficient implementation of the automatic generic transdimensional algorithm of Green (2003), which uses our preferred marginalisation. The efficiency of these methods is compared with several previously proposed samplers on some gene expression data sets. The samplers considered are: the data augmentation method of Holmes and Held (2006), the automatic generic transdimensional algorithm of Green (2003) and the efficient jump proposal methods of Brooks *et al* (2003). Finally, the results of these applications lead us to propose guidelines for choosing between samplers.

Keywords: Probit model, Bayesian variable selection, Data augmentation, Transdimensional Markov chain, Reversible jump sampler, Gene expression data .

*Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. and † Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, U.K.

1 Introduction

In many areas of statistics, we are interested in identifying covariates that discriminate between two classes. For example, in a gene expression experiment it is common to measure the expression level of many genes for a few tissue samples, such as diseased or non-diseased. Only a subset of the genes are needed to successfully discriminate the different states. The goal of a statistical analysis of these data is to identify this small subset of genes that are linked to the molecular mechanism underlining the diseases. This is complicated by the large number of potential subsets and high correlation between many expression levels.

There are two possible approaches to this problem. Most variable selection methods in the literature are univariate in the sense that each candidate gene is considered individually. Examples of univariate methods include the signal to noise ratio of Golub *et al* (1999), the t-test of Nguyen and Rocke (2002) and the ratio of between-groups to within-groups sum of squares of Dudoit *et al* (2002). It is necessary to adjust the nominal significance levels of tests to account for multiple testing. Alternatively we can model class membership as a binary regression on gene expression levels. The statistical problem becomes one of variable selection in a binary regression model. Often, a Bayesian approach is adopted to identify a subset of relevant genes that can give good classification rules. This approach considers multiple genes simultaneously and, hence, naturally accounts for dependence between genes. However the standard Bayesian approach to model selection described by, amongst others, Chipman *et al* (2001) encounters two related problems when applied to the variable selection problem in the probit model with many explanatory variables. Firstly the marginal likelihood for each possible model is not available in analytic form and secondly the number of candidate models is very large, prohibiting the calculation of the posterior model distribution.

There are at least two different approaches that address these problems. In the first approach we efficiently identify a reduced set of good models and use an approximation to compute the marginal likelihood for each possible model. Yeung *et al* (2005) used both the leaps and bounds algorithm and Occam's window to identify a set of good models with a logit link. They approximated the marginal likelihood for each model with the Bayesian information criterion (BIC). Hans *et al* (2007) introduced a shotgun stochastic search method that uses parallel computing to evaluate and record many good models. The marginal likelihood is ap-

proximated by the Laplace method. The second approach applies Markov chain Monte Carlo methodology that simultaneously explores the model and parameter space. The class of Markov chains that admit transitions between states of differing dimension are termed transdimensional Markov chains. A comprehensive survey can be found in Sisson (2005). We will concentrate on developing and implementing transdimensional Markov chains that are special forms of the reversible jump sampler introduced by Green (1995). For example, Holmes and Held (2006), Sha *et al* (2004) and Lee *et al* (2003) used the data augmentation approach described by Albert and Chib (1993) to define efficient reversible jump samplers. In this case the reversible jump acceptance probability is independent of both current and proposed parameter states. Therefore the reversible jump sampler becomes a fixed-dimensional one over the space of models. However the data augmentation approach can cause slow mixing in the chain since the auxiliary variables are correlated with the model and the model parameters. In this paper we avoid this problem by applying existing forms of reversible jump sampler that jointly update the model and the auxiliary variables. Therefore the auxiliary variables are not used when updating the model. The first one is the automatic generic transdimensional sampler proposed by Green (2003), which uses an approximation to the posterior distribution to aid mixing. We consider the Laplace approximation and the modified Iterative Weighted Least Square method described by Gamerman (1997), which can lead to reduced CPU times. The other algorithms that we apply are the higher order and conditional maximization methods introduced by Brooks *et al* (2003) to achieve the automatic scaling and location of the proposal density in reversible jump samplers.

A second aim of this paper is the extension of the local model proposal implemented by Sha *et al* (2004) to a more general one. The model proposal is an important component of transdimensional algorithms. In our experience, a model proposal that randomly chooses to either add or delete a single explanatory variable or to swap two explanatory variables in the current model often leads to high model acceptance rates when applied to problems with many more variables than observations. We consider generalizing this model proposal by adding, deleting or swapping several variables. This should lead to better mixing since a Metropolis random walk with local proposals and high acceptance rate is often associated with poor mixing. More global moves updating a block of explanatory variables leads to model updates with lower acceptance rate but better mixing.

Finally, the efficiency and mixing performance of all transdimensional algorithms

described in this paper are evaluated and compared using some gene expression datasets. The main findings of these comparisons lead us to propose guidelines that optimize MCMC efficiency.

2 The Bayesian Model

Suppose that we observe responses $\boldsymbol{y} = (y_1, \ldots y_n)'$ taking the values 0 or 1 which indicates class membership. The probit model assumes that the probability $\pi(y_i = 1) = p_i$ is modelled by

$$y_i|p_i \sim \text{Bernoulli}(p_i = \Phi(\eta_i))$$
$$\boldsymbol{\eta} = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X}$ is an $n \times p$ matrix whose (i, j)-th entry is the measurement of the j-th covariate for the i-th individual, Φ is the cumulative distribution function of a standard normal random variable, $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)$ is a vector of linear predictors, $\mathbf{1}$ represents a $n \times 1$-dimensional vector of ones, α is the intercept and $\boldsymbol{\beta}$ represents a $p \times 1$-dimensional vector of regression coefficients. We assume that the covariates have been centred.

In the variable selection problem for the probit model we aim to model the relationship between the response $\boldsymbol{y}$ and a (small) subset of the p explanatory variables. There are 2^p possible subset choices and for convenience these are indexed by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ where $\gamma_i = 0$ or 1 according to whether the i-th predictor is excluded from or included in the model. The number of variables included in a model is denoted by $p_\gamma = \sum_{i=1}^{p} \gamma_i$. In line with the bulk of the literature for variable selection with linear regression models, see for example Mitchell and Beauchamp (1988) and Brown *et al* (1998a), exclusion of a variable means that the corresponding element of $\boldsymbol{\beta}$ is zero. Thus, a model indexed by γ containing p_γ variables is defined by

$$y_i|\alpha, \boldsymbol{\beta}_\gamma, \mathbf{x}_{\gamma i} \sim \text{Bernoulli}(p_i = \Phi(\eta_i))$$
$$\boldsymbol{\eta} = \alpha\mathbf{1} + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma$$

where $\mathbf{X}_\gamma$ is a $n \times p_\gamma$ matrix whose columns are the included variables and $\boldsymbol{\beta}_\gamma$ is a $p_\gamma \times 1$-dimensional vector of regression coefficients. We denote the model parameters by $\boldsymbol{\theta}_\gamma = (\alpha, \boldsymbol{\beta}_\gamma')' \in \boldsymbol{\Theta}_\gamma$.

The Bayesian approach specifies a prior distribution for the intercept α, the regression coefficients $\boldsymbol{\beta}_\gamma$ and the model γ which usually has the following structure

$$\pi(\alpha, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})\pi(\alpha)\pi(\boldsymbol{\gamma}).$$

The prior distribution for the regression coefficients $\boldsymbol{\beta}_\gamma$ is given by

$$\pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) \quad \sim \quad N_{p_\gamma}(\mathbf{0}, \boldsymbol{V}_\gamma) \tag{2.1}$$

where $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a p-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We will assume that $\boldsymbol{V}_\gamma$ is a diagonal matrix $c\mathbf{I}_{p_\gamma}$ which yields the ridge prior used by Denison *et al* (2002). This implies that the coefficients are independent *a priori*. Alternatively, a g-prior where $\boldsymbol{V}_\gamma = c(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}$ could be used. Turning to the intercept α, Sha *et al* (2004) and Brown *et al* (1998a) have used a univariate normal $N(0, h)$, where h is large, and this is the one we adopt here. The regressors have been centred and so α represents the overall mean of the linear predictors and it is regarded as a common parameter to all models. As a consequence the non-informative improper uniform prior for location parameters can also be used. We assume that each regressor is included in the model independently with probability w which implies that

$$\pi(\gamma) = w^{p_\gamma}(1-w)^{p-p_\gamma} \tag{2.2}$$

and p_γ is binomially distributed $\text{Bin}(p, w)$. Therefore the model size has prior mean pw and variance $pw(1-w)$. Increased prior probability on parsimonious models could be obtained by setting w small.

This Bayesian approach to variable selection for the probit model accounts for dependency between explanatory variables and simpler models are favored over more complex ones when comparable fits are provided to the data. Therefore, a small subset of relevant explanatory variables is expected to be selected. The choice of the hyperparameters w and c is quite critical for the posterior inference of Bayesian variable selection since w plays the main role in inducing a size penalty and c is inducing regulation on the regression coefficients. There is no clear evidence of a trade-off between w and c in probit regression with $p >> n$, in contrast with the trade-off described in Ley and Steel (2007) for the linear regression model.

3 Posterior Inference and Exploration

Posterior inference using this prior for the probit model is complicated by the lack of an analytic form of the marginal likelihood $\pi(\boldsymbol{y}|\boldsymbol{\gamma})$ of model $\boldsymbol{\gamma}$. Consequently, we either approximate the marginal likelihood allowing us to define an approximate posterior distribution on model space which can be searched directly by Metropolis-Hasting sampling or we run an MCMC sampler on the joint space $(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})$. Here we shall avoid approximations and use the latter approach. A second problem in our case is the large number of candidate models due to the large number of explanatory variables.

To sample the model and model parameters jointly we will construct a Markov chain with state space $\boldsymbol{\Theta} = \bigcup_\gamma \boldsymbol{\Theta}_\gamma \times \{\boldsymbol{\gamma}\}$ and stationary distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y})$. The state space $\boldsymbol{\Theta}$ is a finite union of subspaces of varying dimension and the stationary distribution π is absolute continuous in $\boldsymbol{\theta}_\gamma$ for each $\boldsymbol{\gamma}$ with respect to $(p_\gamma + 1)$-dimensional Lebesque measure and can be sampled using reversible jump Metropolis-Hastings (Green 1995).

Posterior simulation of the probit model can be greatly helped by the data augmentation approach of Albert and Chib (1993). Auxiliary variables $z_1, \ldots, z_n$ are introduced such that

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise.} \end{cases} \\ \boldsymbol{z} &= \tilde{\mathbf{X}}_\gamma \boldsymbol{\theta}_\gamma + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N_n(\mathbf{0}, \mathbf{I}_n) \end{aligned} \tag{3.3}$$

where y_i is now deterministic conditional on the sign of the stochastic auxiliary variable z_i and $\tilde{\mathbf{X}}_\gamma = (\mathbf{1} : \mathbf{X}_\gamma)$ is the design matrix corresponding to model γ. The full conditional distribution can then be sampled directly (z_i is truncated normal and $\boldsymbol{\theta}_\gamma$ is multivariate normal).

Sha *et al* (2004) used the data augmentation approach and integrated out the model parameters $\boldsymbol{\theta}_\gamma$. The target distribution of their sampler is the joint posterior distribution $\pi(\boldsymbol{z}, \boldsymbol{\gamma}|\boldsymbol{y})$. They used the Metropolis-Hastings algorithm to sample $\boldsymbol{\gamma}$ conditional on $\boldsymbol{z}$ and then sampled $\boldsymbol{z}$ from its full conditional distribution $\boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{y}$ which is multivariate truncated normal and can be sampled using the sub chain Gibbs sampler of Geweke (1991).

Alternatively, we could define a Gibbs sampler for $\boldsymbol{z}, \boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}$. Samplers that update each parameter individually may have mixing problems and we consider jointly

updating some parameters with the model. The algorithm defined by Holmes and Held (2006) updates $\boldsymbol{\gamma}, \boldsymbol{\theta}_\gamma$ jointly. The full conditional distribution can be expressed as $p(\boldsymbol{\gamma}|\boldsymbol{z})p(\boldsymbol{\theta}_\gamma|\boldsymbol{z}, \boldsymbol{\gamma})$. Alternatively, the Automatic Generic and Efficient Proposal samplers update $\boldsymbol{\gamma}, \boldsymbol{z}$ jointly by updating $\boldsymbol{\gamma}$ given $\boldsymbol{\theta}_\gamma$ and $\boldsymbol{z}$ given $\boldsymbol{\gamma}, \boldsymbol{\theta}_\gamma$. In each case, all other parameters are updated using Gibbs sampler updates. All samplers have common update steps for $\boldsymbol{z}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}, \boldsymbol{y}$ and $\boldsymbol{\theta}_\gamma|\boldsymbol{z}, \boldsymbol{\gamma}$. These steps are standard and have the pseudo-code:

1. Update $\boldsymbol{z}$ from its full conditional distribution $\boldsymbol{z}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}, \boldsymbol{y}$. The full conditional of z_i is a normal distribution with mean $\alpha + \mathbf{x}_i\boldsymbol{\beta}_\gamma$ and variance 1 truncated to $(0, \infty)$ if $y_i = 1$ or $(-\infty, 0)$ otherwise. These distributions can be efficiently sampled using the optimized exponential rejection sampling method described by Geweke (1991).

2. Update the parameter vector $\boldsymbol{\theta}_\gamma$ from its full conditional distribution $\boldsymbol{\theta}_\gamma|\boldsymbol{z}, \boldsymbol{\gamma}$. This is a multivariate normal given by

$$\begin{aligned} \boldsymbol{\theta}_\gamma|\boldsymbol{z}, \boldsymbol{\gamma} &\sim N_{p_\gamma+1}\left((\tilde{\mathbf{X}}_\gamma'\tilde{\mathbf{X}}_\gamma + \mathbf{H}_\gamma^{-1})^{-1}\tilde{\mathbf{X}}_\gamma'\boldsymbol{z}, (\tilde{\mathbf{X}}_\gamma'\tilde{\mathbf{X}}_\gamma + \mathbf{H}_\gamma^{-1})^{-1}\right) \\ \mathbf{H}_\gamma &= \begin{bmatrix} h & \mathbf{0} \\ \mathbf{0} & c\mathbf{I}_{p_\gamma} \end{bmatrix}. \end{aligned} \tag{3.4}$$

   In the case of the improper uniform prior on α we obtain a very similar full conditional.

3.1 Between-model moves

The model space has a varying dimension and updating will make use of reversible jump Metropolis-Hastings methods (Green 1995). A new parameter vector $\boldsymbol{\theta}_{\gamma'}$ for model $\boldsymbol{\gamma'}$ is proposed using both the current parameter vector $\boldsymbol{\theta}_\gamma$ of model $\boldsymbol{\gamma}$ and a random vector. The standard Metropolis-Hastings acceptance probability is also modified to account for the varying dimension of the state space. The idea is to supplement each of the spaces $\boldsymbol{\Theta}_\gamma$ and $\boldsymbol{\Theta}_{\gamma'}$ with adequate artificial spaces in order to create a bijection map between them. We are going to describe the reversible jump sampler in the Bayesian variable selection setting.

We assume that the current state of the Markov chain is $(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})$ and the model proposal $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$ generates the new model $\boldsymbol{\gamma'}$. If the current model parameter $\boldsymbol{\theta}_\gamma$ is completed by a random variable $\boldsymbol{u}_\gamma \sim q_\gamma(\boldsymbol{u})$ into $(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma)$, and $\boldsymbol{\theta}_{\gamma'}$ by $\boldsymbol{u}_{\gamma'} \sim q_{\gamma'}(\boldsymbol{u})$ into $(\boldsymbol{\theta}_{\gamma'}, \boldsymbol{u}_{\gamma'})$ so that the map $(\boldsymbol{\theta}_{\gamma'}, \boldsymbol{u}_{\gamma'}) = g(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma)$ is bijective then the probability of acceptance for the move from model $\boldsymbol{\gamma}$ to model $\boldsymbol{\gamma'}$ is $\min\{1, A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \to (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]\}$. Here

$$A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \to (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})] = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})\, \pi(\boldsymbol{\theta}_{\gamma'}|\boldsymbol{\gamma'})\, \pi(\boldsymbol{\gamma'})\, q_{\gamma'}(\boldsymbol{u}_{\gamma'})\, q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})}{\pi(\boldsymbol{y}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})\, \pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma})\, \pi(\boldsymbol{\gamma})\, q_\gamma(\boldsymbol{u}_\gamma)\, q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})} \left| \frac{\partial g(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma)}{\partial(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma)} \right|, \tag{3.5}$$

involving the Jacobian of the transform g, the probability $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$ of proposing to move from model $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma'}$ and q_γ which is the density of $\boldsymbol{u}_\gamma$. This proposal satisfies the detailed balance condition and the symmetry assumption of Green (1995). The stationary distribution of this Markov chain is the joint posterior distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y})$. The pseudo-code representation of Green's algorithm is as follow:
If at iteration t the current state is $(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{\gamma})$ then

1. Select model $\boldsymbol{\gamma'}$ with probability $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$.

2. Generate $\boldsymbol{u}_\gamma \sim q_\gamma(\boldsymbol{u})$.

3. Set $(\boldsymbol{\theta}_{\gamma'}, \boldsymbol{u}_{\gamma'}) = g(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{u}_\gamma)$'

4. Jump to the model $\boldsymbol{\gamma'}$ and set $\boldsymbol{\theta}_{\gamma'}^{(t+1)} = \boldsymbol{\theta}_{\gamma'}$ with probability

$$\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma'}) = \min\{1, A[(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{\gamma}) \to (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]\}$$

otherwise take $\boldsymbol{\theta}_\gamma^{(t+1)} = \boldsymbol{\theta}_\gamma^{(t)}$.
Here $A[(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{\gamma}) \to (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]$ is given by (3.5).

3.1.1 Holmes and Held algorithm

Holmes and Held (2006) and Lee *et al* (2003) choose a proposal that reduces the reversible jump sampler to a fixed-dimensional one over the space of models. If the random vector $\boldsymbol{u}_\gamma \sim q_\gamma(\boldsymbol{u}) = \pi(\boldsymbol{\theta}_{\gamma'}|\boldsymbol{\gamma'}, \boldsymbol{z})$ is a draw directly from its conditional distribution and the proposal state $\boldsymbol{\theta}_{\gamma'} = \boldsymbol{u}_\gamma$ then the acceptance probability (3.5) reduces to

$$A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \to (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})] = \frac{\pi(\boldsymbol{\gamma'})\, q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})\, \pi(\boldsymbol{z}|\boldsymbol{\gamma'})}{\pi(\boldsymbol{\gamma})\, q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})\, \pi(\boldsymbol{z}|\boldsymbol{\gamma})} \tag{3.6}$$

The above acceptance probability is independent of both current and proposed parameter states and it is similar to the acceptance probability of Metropolis-Hastings algorithm with target distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{z})$. Thereby the reversible jump sampler becomes a fixed dimensional one over the space of models. The pseudo-code of Holmes and Held algorithm proceeds as follow:
If at iteration t the current state is $(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}_{\gamma}^{(t)}, \boldsymbol{\gamma})$ then

1. Select model $\boldsymbol{\gamma'}$ with probability $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$.

2. Jump to the model $\boldsymbol{\gamma'}$ with probability

$$\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma'}) = \min\{1, A[\boldsymbol{\gamma} \rightarrow \boldsymbol{\gamma'}]\}$$

   Here $A[\boldsymbol{\gamma} \rightarrow \boldsymbol{\gamma'}]$ is given by (3.6).

3. If the jump to model $\boldsymbol{\gamma'}$ is accepted draw a sample $\boldsymbol{\theta}_{\gamma'} \sim \pi(\boldsymbol{\theta}_{\gamma'}|\boldsymbol{\gamma'}, \boldsymbol{z}^{(t)})$ and set $\boldsymbol{\theta}_{\gamma'}^{(t+1)} = \boldsymbol{\theta}_{\gamma'}$. Otherwise set $\boldsymbol{\theta}_{\gamma}^{(t+1)} = \boldsymbol{\theta}_{\gamma}^{(t)}$.

The Holmes and Held sampler is likely to mix slowly because the auxiliary variable $\boldsymbol{z}$ is correlated with $(\boldsymbol{\theta}_{\gamma}, \boldsymbol{\gamma})$, as it is seen from (3.3), and a Gibbs sampler is used to update $\boldsymbol{z}$. Similarly the Sha *et al* (2004) sampler may face the same problem since $\boldsymbol{z}$ is correlated with $\boldsymbol{\gamma}$ and a Gibbs sampler is used to update $\boldsymbol{z}$.

3.1.2 Automatic Generic Sampler

This algorithm was introduced by Green (2003) and reparameterizes from $\boldsymbol{\theta}_{\gamma}$ to $\boldsymbol{\nu}$ where

$$\boldsymbol{\theta}_{\gamma} = \boldsymbol{\mu}_{\gamma} + \boldsymbol{B}_{\gamma}\boldsymbol{\nu}$$

where $\boldsymbol{\mu}_{\gamma}$ approximates the mean of $\pi(\boldsymbol{\theta}_{\gamma}|\boldsymbol{\gamma}, \boldsymbol{y})$ and $\boldsymbol{B}_{\gamma}$ approximates the Cholesky factor of the covariance matrix of $\pi(\boldsymbol{\theta}_{\gamma}|\boldsymbol{\gamma}, \boldsymbol{y})$. Proposing a new model $\boldsymbol{\gamma'}$ then we set a new vector $\boldsymbol{\theta}_{\gamma'}$ to be:

$$\boldsymbol{\theta}_{\gamma'} = \begin{cases} \boldsymbol{\mu}_{\gamma'} + \boldsymbol{B}_{\gamma'}\left(\boldsymbol{R}\boldsymbol{B}_{\gamma}^{-1}(\boldsymbol{\theta}_{\gamma} - \boldsymbol{\mu}_{\gamma})\right)_1^{p_{\gamma'}} & \text{if} \quad p_{\gamma'} < p_{\gamma} \\ \boldsymbol{\mu}_{\gamma'} + \boldsymbol{B}_{\gamma'}\boldsymbol{R}\boldsymbol{B}_{\gamma}^{-1}\left(\boldsymbol{\theta}_{\gamma} - \boldsymbol{\mu}_{\gamma}\right) & \text{if} \quad p_{\gamma'} = p_{\gamma} \\ \boldsymbol{\mu}_{\gamma'} + \boldsymbol{B}_{\gamma'}\boldsymbol{R}\begin{bmatrix} \boldsymbol{B}_{\gamma}^{-1}(\boldsymbol{\theta}_{\gamma} - \boldsymbol{\mu}_{\gamma}) \\ \boldsymbol{u}_{\gamma} \end{bmatrix} & \text{if} \quad p_{\gamma'} > p_{\gamma} \end{cases} \tag{3.7}$$

Here $(\cdot)_1^m$ denotes the first m component of a vector, $\boldsymbol{R}$ is a fixed orthogonal matrix of order $\max\{p_{\gamma}, p_{\gamma'}\}$ and $\boldsymbol{u}_{\gamma} \sim q_{\gamma}(\boldsymbol{u})$ is a multivariate random variable of dimension

$(p_{\gamma'} - p_\gamma)$. If $p_{\gamma'} \leq p_\gamma$, then the proposal is deterministic. Since everything is linear, the Jacobian of the transformation is easily calculated and if $p_{\gamma'} > p_\gamma$, we have:

$$\left| \frac{\partial \boldsymbol{\theta}_{\gamma'}}{\partial(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma)} \right| = \frac{|\boldsymbol{B}_{\gamma'}|}{|\boldsymbol{B}_\gamma|}.$$

Thus the acceptance probability of moving to model $\boldsymbol{\gamma'}$ is $\min\{1, A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]\}$ and (3.5) takes the form

$$A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})] = \frac{\pi(\boldsymbol{\gamma'}, \boldsymbol{\theta}_{\gamma'}|\boldsymbol{y})\ q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})\ |\boldsymbol{B}_{\gamma'}|}{\pi(\boldsymbol{\gamma}, \boldsymbol{\theta}_\gamma|\boldsymbol{y})\ q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})\ |\boldsymbol{B}_\gamma|} \times \begin{cases} q_\gamma(\boldsymbol{u}_\gamma) & \text{if} \quad p_{\gamma'} < p_\gamma \\ 1 & \text{if} \quad p_{\gamma'} = p_\gamma \\ q_\gamma(\boldsymbol{u}_\gamma)^{-1} & \text{if} \quad p_{\gamma'} > p_\gamma \end{cases} \tag{3.8}$$

Since $\boldsymbol{R}$ is orthogonal it does not play any role in this calculation. The author's motivation in developing this algorithm is that high transition probabilities may be achieved when $\pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{y})$ are reasonably unimodal and the first and second moments are approximately equal to $\boldsymbol{\mu}_\gamma$ and $\boldsymbol{B}_\gamma \boldsymbol{B}'_\gamma$. The pseudo-code representation of this sampler is as follow:
If at iteration t the current state is $(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{\gamma})$ then

1. Select model $\boldsymbol{\gamma'}$ with probability $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$.
2. Generate $\boldsymbol{u}_\gamma \sim q_\gamma(\boldsymbol{u})$.
3. Set the new parameter vector $\boldsymbol{\theta}_{\gamma'}$ using (3.7).
4. Jump to the model $\boldsymbol{\gamma'}$ and set $\boldsymbol{\theta}_{\gamma'}^{(t+1)} = \boldsymbol{\theta}_{\gamma'}$ with probability

$$\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma'}) = \min\{1, A[(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]\}$$

   otherwise take $\boldsymbol{\theta}_\gamma^{(t+1)} = \boldsymbol{\theta}_\gamma^{(t)}$.
   Here $A[(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]$ is given by (3.8).

We consider two methods to approximate the first and second moments of $\pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{y})$. The first is the Laplace method and the second is a Bayesian version of the Iterative Weighted Least Square algorithm described by Gamerman (1997). The Laplace method approximates the mean and covariance matrix of $\pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{y})$ by its posterior mode $\hat{\boldsymbol{\mu}}_\gamma$ and the matrix $\hat{\boldsymbol{\Sigma}}_\gamma$, the inverse of the negative Hessian matrix at $\hat{\boldsymbol{\mu}}_\gamma$, respectively. This method solves an optimization problem in each iteration and therefore is computationally not efficient.

The automatic generic sampler can propose reasonable values of $\boldsymbol{\theta}_{\gamma'}$ and achieve high acceptance rate even when the estimates of the first and second moments are not very accurate. We use the Bayesian Iterative Weighted Least Square algorithm (Gamerman 1997) to find rough estimates of the first and second moments. This algorithm finds the posterior mode $\hat{\boldsymbol{\mu}}_\gamma$ by iterating

$$\boldsymbol{\mu}_\gamma^{(t)} = \left(\mathbf{H}_\gamma^{-1} + \tilde{\mathbf{X}}_\gamma' \boldsymbol{W}\left(\boldsymbol{\mu}_\gamma^{(t-1)}\right)\tilde{\mathbf{X}}_\gamma\right)^{-1} \tilde{\mathbf{X}}_\gamma' \boldsymbol{W}\left(\boldsymbol{\mu}_\gamma^{(t-1)}\right)\tilde{\boldsymbol{y}}\left(\boldsymbol{\mu}_\gamma^{(t-1)}\right)$$

until convergence, where $\mathbf{H}_\gamma$ is the prior covariance matrix of the intercept and the regression coefficients given by (3.4), $\tilde{\mathbf{X}}_\gamma = (\mathbf{1} : \mathbf{X}_\gamma)$ is the design matrix corresponding to model $\boldsymbol{\gamma}$, $\tilde{\boldsymbol{y}}\left(\boldsymbol{\mu}_\gamma^{(t-1)}\right)$ is a vector of transformed observations and $\boldsymbol{W}\left(\boldsymbol{\mu}_\gamma^{(t-1)}\right)$ is a diagonal matrix of weights. The inverse of curvature at $\hat{\boldsymbol{\mu}}_\gamma$ is given by $\left(\mathbf{H}_\gamma^{-1} + \tilde{\mathbf{X}}_\gamma \boldsymbol{W}\left(\hat{\boldsymbol{\mu}}_\gamma\right)\tilde{\mathbf{X}}_\gamma\right)^{-1}$. In the case of the binary probit model the vector of transformed observations $\tilde{\boldsymbol{y}}\left(\boldsymbol{\mu}_\gamma^{(t-1)}\right)$ is defined as

$$\tilde{y}_i\left(\boldsymbol{\mu}_\gamma^{(t-1)}\right) = \tilde{\mathbf{x}}_{\gamma i}\boldsymbol{\mu}_\gamma^{(t-1)} + (y_i - \mathbf{E}(y_i))\frac{d\eta_i}{dp_i} = \eta_i + (y_i - p_i)\frac{1}{\phi(\eta_i)} \qquad i = 1, \ldots, n$$

and the diagonal matrix of weights $\boldsymbol{W}\left(\boldsymbol{\mu}_\gamma^{(t-1)}\right)$ is defined as:

$$w_{ii} = \frac{1}{\mathrm{Var}(y_i)}\left(\frac{dp_i}{d\eta_i}\right)^2 = \frac{1}{p_i(1-p_i)}\phi(\eta_i)^2 = \frac{\phi(\eta_i)^2}{\Phi(\eta_i)(1-\Phi(\eta_i))} \qquad i = 1, \ldots, n$$

where $\eta_i = \tilde{\mathbf{x}}_{\gamma i}\boldsymbol{\mu}_\gamma^{(t-1)}$, $\tilde{\mathbf{x}}_{\gamma i}$ is the ith row of the design matrix $\tilde{\mathbf{X}}_\gamma$, ϕ is the probability density function of the standard normal and Φ is the corresponding cumulative distribution function. We propose to use either a single or more iteration cycles of this modified Iterative Weighted Least Square method to find rough estimates of the first and second moments of $\pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{y})$. Thus, this method is computationally more efficient than the Laplace approximation.

3.1.3 Efficient Construction of Reversible Jump Proposal Densities

Brooks *et al* (2003) discusses a collection of techniques that can be used to scale and shape automatically the reversible jump proposal distribution $q_\gamma(\boldsymbol{u})$. The proposal parameters are adapted to the current state of the chain at each stage, rather than relying on a constant proposal parameter vector for all state transitions. This group of methods is based on an analysis of acceptance probability for jumps which involves

Taylor series expansion of the acceptance probability (3.5) around certain canonical jumps.

In what follows we assume that the current state of the chain is $\boldsymbol{\theta}_\gamma \in \boldsymbol{\Theta}_\gamma$ and we propose to move to model $\boldsymbol{\gamma'}$ using the model proposal $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$. Brooks *et al* (2003) focus on moves between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma'}$ such that $\dim(\boldsymbol{\Theta}_{\gamma'}) > \dim(\boldsymbol{\Theta}_\gamma)$. By reversibility, this also characterizes the reverse move. Between each collection of models for which they might attempt to jump they fix the between model mapping $g(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma(\boldsymbol{\upsilon}))$, where $\boldsymbol{u}_\gamma$ is a general proposal transformation of some canonical random $\boldsymbol{\upsilon}$. They define the centering function $c : \boldsymbol{\Theta}_\gamma \to \boldsymbol{\Theta}_{\gamma'}$ by the equation

$$c(\boldsymbol{\theta}_\gamma) = g(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma)))$$

where $\boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma))$ is a specific value for the proposal vector $\boldsymbol{u}_\gamma$. Equivalently $\boldsymbol{b}(\boldsymbol{\theta}_\gamma)$ is a specific value for the canonical random vector $\boldsymbol{\upsilon}$. They propose to specify this particular value $\boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma))$ such that, the current value $\boldsymbol{\theta}_\gamma$ and the $c(\boldsymbol{\theta}_\gamma)$ are identical in terms of likelihood contribution: that is $\pi(\boldsymbol{y}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) = \pi(\boldsymbol{y}|c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})$.

In our application the reversible jump proposal is $\boldsymbol{u}_\gamma(\boldsymbol{\upsilon}) = \boldsymbol{\mu} + \sigma\boldsymbol{\upsilon}$ which is a linear transformation of $\boldsymbol{\upsilon}$ and $\boldsymbol{\upsilon} \sim N_{p_{\gamma'}-p_\gamma}(\mathbf{0}, \mathbf{I}_{p_{\gamma'}-p_\gamma})$, that is the standard multivariate normal of dimension $p_{\gamma'} - p_\gamma$. The between-model map is set to the identity, that is $g(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma(\boldsymbol{\upsilon})) = (\boldsymbol{\theta}_\gamma, \boldsymbol{\mu} + \sigma\boldsymbol{\upsilon})$. Therefore the centering function for a move between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma'}$ for the variable selection problem is $c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, \mathbf{0})$ since the $(p_\gamma + 1)$-dimensional model with parameter vector $\boldsymbol{\theta}_\gamma$ is identical in terms of likelihood contribution with the $(p_{\gamma'} + 1)$-dimensional model with parameters $(\boldsymbol{\theta}_\gamma, \mathbf{0})$. Thus the likelihood drops out of equation (3.5) since $\pi(\boldsymbol{y}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) = \pi(\boldsymbol{y}|c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})$. Furthermore the Jacobian term of (3.5) is

$$\left| \frac{\partial(\boldsymbol{\theta}_\gamma, \boldsymbol{\mu} + \sigma\boldsymbol{\upsilon})}{\partial(\boldsymbol{\theta}_\gamma, \boldsymbol{\upsilon})} \right| = \sigma^{p_{\gamma'}-p_\gamma}. \tag{3.9}$$

Brooks *et al* (2003) introduced general methods to obtain the location $\boldsymbol{\mu}$ and the scale σ of the proposal random variable $\boldsymbol{u}_\gamma$ and we will show how to implement them in the variable selection problem for the probit model. These methods differ in the order of the Taylor series expansion of (3.5) around the centering point $c(\boldsymbol{\theta}_\gamma)$.

Zeroth Order method

This method automatically specifies the scale of the proposal transformation $\boldsymbol{u}_\gamma(\boldsymbol{\upsilon}) = \sigma\boldsymbol{\upsilon}$, where the location parameter $\boldsymbol{\mu}$ is assumed to be $\mathbf{0}$. The scale is

chosen so that, for the jump between $\boldsymbol{\theta}_\gamma$ and its image in $\boldsymbol{\theta}_{\gamma'}$ under the centering function $c(\boldsymbol{\theta}_\gamma)$, the acceptance ratio (3.5) equals 1, that is

$$A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow (c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})] = 1. \tag{3.10}$$

The following mathematical relations hold

$$c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, \mathbf{0}) \quad \Leftrightarrow \quad \boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma)) = \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{b}(\boldsymbol{\theta}_\gamma) = \mathbf{0}. \tag{3.11}$$

If we substitute both (3.9) and (3.11) into (3.10) and rearrange we obtain

$$\sigma = \left(\frac{c^{(p_{\gamma'} - p_\gamma)/2} \pi(\boldsymbol{\gamma}) q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma'}) q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})} \right)^{\frac{1}{p_{\gamma'} - p_\gamma}} \tag{3.12}$$

where c is the hyperparameter that determines the prior covariance matrix $\mathbf{V}_\gamma = c\mathbf{I}_{p_\gamma}$ of the regression coefficients $\boldsymbol{\beta}_\gamma$.

Higher Order methods

The proposal variance using the zeroth order method is independent of the data and so only information from the prior is used to tune the proposal distribution. The method may be improved if we can also incorporate information from the data in choosing the proposal scale. A natural way to do this is to consider higher order approximations that require the first r derivatives of the logarithm of the acceptance probability to equal the zero vector at $c(\boldsymbol{\theta}_\gamma)$, that is

$$\nabla^r \log A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow (c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})] = \mathbf{0}.$$

Here the partial derivatives are taken with respect to $\boldsymbol{\upsilon}$. As we set increasingly more derivatives to $\mathbf{0}$ we obtain acceptance probabilities closer to 1, at least in some neighbourhood of the centering point $c(\boldsymbol{\theta}_\gamma)$. In practise our proposal density will typically have few parameters which need to be selected. Given a proposal with κ parameters we only need κ constraints to specify those parameters. The first order method satisfies the system of equations

$$\begin{aligned} A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow (c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})] &= 1 \\ \nabla^1 \log A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow (c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})] &= \mathbf{0} \end{aligned}$$

for all possible choices of $\boldsymbol{\gamma}, \boldsymbol{\gamma'}$ and $\boldsymbol{\theta}_\gamma$, which imposes an $p_{\gamma'} - p_\gamma + 1$ dimensional constraint on the proposal. The location and scale of the proposal transformation

$\boldsymbol{u}_\gamma(\boldsymbol{v}) = \boldsymbol{\mu} + \sigma\boldsymbol{v}$ are the solutions to the above system of equations. The following mathematical relations hold

$$c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, \mathbf{0}) \quad \Leftrightarrow \quad \boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma)) = \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{b}(\boldsymbol{\theta}_\gamma) = -\frac{\boldsymbol{\mu}}{\sigma}.$$

The system of equations is written as

$$\begin{aligned} 1 &= \frac{\pi(\boldsymbol{\gamma'})q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})}{\pi(\boldsymbol{\gamma})q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})\exp -\frac{\boldsymbol{\mu'\mu}}{2\sigma^2}} \left(\frac{\sigma}{c}\right)^{p_{\gamma'}-p_\gamma} \\ \tfrac{1}{\sigma^2}\boldsymbol{\mu} &= \mathbf{X}_{p_{\gamma'}-p_\gamma}\mathbf{D}_1\boldsymbol{y} - \mathbf{X}_{p_{\gamma'}-p_\gamma}\mathbf{D}_2(\mathbf{1}-\boldsymbol{y}) \end{aligned}$$

where $\mathbf{X}_{p_{\gamma'}-p_\gamma}$ is a $(p_{\gamma'} - p_\gamma) \times n$ matrix with entries the measurements of the new variables proposed to be included, $\mathbf{D}_1$ is a diagonal matrix with elements $(\frac{\phi(\eta_1)}{\Phi(\eta_1)}, \ldots, \frac{\phi(\eta_n)}{\Phi(\eta_n)})$, $\mathbf{D}_2$ is a diagonal matrix with elements $(\frac{\phi(\eta_1)}{\Phi(-\eta_1)}, \ldots, \frac{\phi(\eta_n)}{\Phi(-\eta_n)})$ and $\boldsymbol{\eta} = \tilde{\mathbf{X}}_\gamma\boldsymbol{\theta}_\gamma$. This system of equations can not be solved analytically and requires a numerical solution which is computationally demanding. Since the acceptance ratio is 1 except for a quadratic error, larger jumps can be attempted without leading to acceptance rates close to 0.

The second order method sets the first and second derivatives of the logarithm of the acceptance probability equal to $\mathbf{0}$ at $c(\boldsymbol{\theta}_\gamma)$, that is:

$$\nabla^r \log A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow (c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})] = \mathbf{0} \qquad r = 1, 2. \tag{3.13}$$

There are two drawbacks with this method. Firstly, there could be more constraints than the proposal parameters needed to be determined. Secondly, it is computationally demanding to apply this method to our problem when the proposal model $\boldsymbol{\gamma'}$ and $\boldsymbol{\gamma}$ differ by more than one explanatory variable that is $p_{\gamma'} > p_\gamma + 1$ since the constraint on the Hessian matrix considerably increases the number of equations to be solved.

If $p_{\gamma'} = p_\gamma + 1$, the system of equations (3.13) involve two constraints and only the two parameters μ and σ need to be determined. The following mathematical relations hold

$$c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, 0) \quad \Leftrightarrow \quad u_\gamma(b(\boldsymbol{\theta}_\gamma)) = 0 \quad \Leftrightarrow \quad b(\boldsymbol{\theta}_\gamma) = -\frac{\mu}{\sigma}.$$

Then the solution to the system of equations (3.13) is given by :

$$\begin{aligned} \sigma^{-2} &= \sum_{i=1}^{n}\left[\frac{y_i\, x_{vi}^2\, \phi(\eta_i)\, (\eta_i\Phi(\eta_i) + \phi(\eta_i))}{(\Phi(\eta_i))^2} + \frac{(1-y_i)\, x_{vi}^2\phi(\eta_i)\, (\phi(\eta_i) - \eta_i\Phi(\eta_i))}{(\Phi(-\eta_i))^2} + \frac{1}{c}\right] \\ \mu &= \sigma^2\sum_{i=1}^{n}\left[\frac{y_i\, x_{vi}\phi(\eta_i)}{\Phi(\eta_i)} - \frac{(1-y_i)\, x_{vi}\phi(\eta_i)}{\Phi(-\eta_i)}\right] \end{aligned}$$

where $\boldsymbol{\eta} = \tilde{\mathbf{X}}_\gamma \boldsymbol{\theta}_\gamma$ and $\mathbf{x}_\upsilon$ is a $1 \times n$-dimensional vector with entries the measurements of the explanatory variable proposed to be included. If $p_{\gamma'} = p_\gamma + 2$ the proposal vector is

$$\boldsymbol{u}_\gamma(\boldsymbol{\upsilon}) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_2 \end{bmatrix} \begin{bmatrix} \upsilon_1 \\ \upsilon_2 \end{bmatrix}$$

in order for the number of proposal parameters to be equal to the number of constraints and to obtain a unique solution.

Conditional Maximization method

The conditional maximization method is also introduced in Brooks *et al* (2003). It proceeds by maximizing the posterior distribution $\pi((\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma)|\boldsymbol{y})$ with respect to $\boldsymbol{u}_\gamma$. The maximizer $\boldsymbol{\mu}$ is the location of the proposal $\boldsymbol{u}_\gamma(\boldsymbol{\upsilon})$ and the centering function for the variable selection problem is $c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, \boldsymbol{\mu})$. Thus, they essentially condition on the current state $\boldsymbol{\theta}_\gamma$ and center at the posterior conditional mode. The scale of the proposal $\boldsymbol{u}_\gamma(\boldsymbol{\upsilon}) = \boldsymbol{\mu} + \sigma\boldsymbol{\upsilon}$ is specified using the centering function $c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, \boldsymbol{\mu})$ and the zeroth order method, so that

$$A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow ((\boldsymbol{\theta}_\gamma, \boldsymbol{\mu}), \boldsymbol{\gamma'})] = 1. \tag{3.14}$$

In order to apply this method to the variable selection problem for the probit model we need to find the maximizer of the following function of u:

$$f(\boldsymbol{u}) = \sum_{i=1}^{n} \left[y_i \ \log \Phi(\tilde{\mathbf{X}}_\gamma \boldsymbol{\theta}_\gamma + \mathbf{X}_u \ \boldsymbol{u}) + (1 - y_i) \ \log \Phi(-(\tilde{\mathbf{X}}_\gamma \boldsymbol{\theta}_\gamma + \mathbf{X}_u \ \boldsymbol{u})) \right] - \frac{\boldsymbol{u}'\boldsymbol{u}}{2c}$$

where $\mathbf{X}_u$ is a $(p_{\gamma'} - p_\gamma) \times n$ matrix with entries the measurements of the new explanatory variables proposed to be included. The following mathematical relations hold

$$c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, \boldsymbol{\mu}) \quad \Leftrightarrow \quad \boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma)) = \boldsymbol{\mu} \quad \Leftrightarrow \quad \boldsymbol{b}(\boldsymbol{\theta}_\gamma) = \mathbf{0}.$$

The scale of the proposal is determined by (3.14) and is given by:

$$\sigma = \left(\frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \ \pi(\boldsymbol{\gamma}) \ q(\boldsymbol{\gamma'}|\boldsymbol{\gamma}) \ c^{(p_{\gamma'} - p_\gamma)/2} \exp \frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{2c}}{\pi(\boldsymbol{y}|(\boldsymbol{\theta}_\gamma, \boldsymbol{\mu}), \boldsymbol{\gamma'}) \pi(\boldsymbol{\gamma'}) q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})} \right)^{1/(p_{\gamma'} - p_\gamma)}.$$

3.2 A new Model Proposal $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$

The model proposal $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$ is an important part of the transdimensional algorithm since it will control convergence of any algorithm. A special class of Metropolis-Hastings algorithms are obtained from the class of model proposals $q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})$ which

are symmetric in $\boldsymbol{\gamma}'$ and $\boldsymbol{\gamma}$. The simplest symmetric transition kernel is

$$q(\boldsymbol{\gamma}'|\boldsymbol{\gamma}) = \frac{1}{p} \quad \text{if} \quad \sum_{i=1}^{p} |\gamma_i' - \gamma_i| = 1 \tag{3.15}$$

Hence the candidate model is generated by randomly changing one component of the current model $\boldsymbol{\gamma}$ and has either one variable more or one variable less than $\boldsymbol{\gamma}$. Madigan and York (1995) used this model proposal in a model selection context to define their MC^3 algorithm. Raftery *et al* (1997) and Fernández *et al* (2001) also used this algorithm for model averaging in linear regression. However this model proposal is not efficient in variable selection problems with large p where we expect parsimonious models to fit the data well. Sha *et al* (2003), for example, in a microarray problem for classification with $p = 999$ genes were expecting that very few genes, around 5, would give good discrimination. In this case the MC^3 algorithm explores the part of the model space which has small model size. Hence, as noted by Hans *et al* (2007), the probability of adding one variable is $(p - p_\gamma)/p$ which is close to 1 since p is large relative to p_γ. Therefore the algorithm spend a large amount of time trying to add a variable before proposing to delete a variable. However the acceptance rate of adding a new variable is equal to the acceptance rate of deleting one variable if our chain is in equilibrium. As a consequence a large number of adding moves are rejected which yields a low between-model acceptance rate.

Brown *et al* (1998b) extended the model proposal (3.15). They proposed to generate a candidate new model $\boldsymbol{\gamma}'$ from the current $\boldsymbol{\gamma}$ by one of two possible moves. The first move is similar to the one used in the MC^3 algorithm. The second move chooses at random one of the currently included variables and at random one of the currently excluded variables. For the new candidate model $\boldsymbol{\gamma}'$ they excluded the previously included variable and included the previously excluded variable. Both Brown *et al* (1998b) and Sha *et al* (2004) applied this model proposal in a variable selection problem for multivariate and probit regression respectively with large p and small n. This model proposal is again not suitable for variable selection with large p because the first type of move is similar to the symmetric kernel (3.15) and therefore yields similar low between models acceptance rate.

Firstly, we decided to split their first move into two moves, the addition and deletion ones, to avoid proposing many more additions than deletions. However, the resulting model proposal only moves locally since the generated model $\boldsymbol{\gamma}'$ differs from current model $\boldsymbol{\gamma}$ by either one or two variables. This local model proposal will

often yield high between-model acceptance rates when applied to problems with many more variables than observations. There are two possible reasons. Firstly, the generated model $\boldsymbol{\gamma}'$ will be similar to the current model $\boldsymbol{\gamma}$ in terms of model fitting since when p is large many explanatory variables are either redundant or highly correlated. Secondly when the sample size n is small the posterior distribution will be relatively flat and the number of models that are well-supported by the data will be large.

Secondly, the high between-model acceptance rate of the local model proposal motivates us to construct a more general model proposal since a Metropolis random walk with local proposal and high acceptance rate is often associated with poor mixing. This new model proposal is able to combine local moves with more global ones by changing simultaneously a block of variables. Thus it is designed to enable the fast exploration of the model space. We first need to determine the maximum number of variables N that we are going to change from the current model $\boldsymbol{\gamma}$. Then at each iteration t of the algorithm we draw a value $N^{(t)}$ from a binomial distribution with parameters $N-1$ and π, that is $N^{(t)} \sim \text{Bin}(N-1, \pi)$ and define three distinct neighbourhood sets of $\boldsymbol{\gamma}$ given by:

- $\boldsymbol{\gamma}^{+}$: This is a set containing neighbouring models of dimension $p_\gamma + (N^{(t)}+1)$ and includes
$$|\boldsymbol{\gamma}^{+}| = \begin{pmatrix} p - p_\gamma \\ N^{(t)}+1 \end{pmatrix}$$
models. The elements of this set are formed by adding $N^{(t)}+1$ new variables to model $\boldsymbol{\gamma}$. The condition $p - p_\gamma \geq N^{(t)}+1$ is always true in our applications since p is large relative to p_γ.

- $\boldsymbol{\gamma}^{-}$: This is a set containing neighbouring models of dimension $p_\gamma - (N^{(t)}+1)$ and includes
$$|\boldsymbol{\gamma}^{-}| = \begin{pmatrix} p_\gamma \\ N^{(t)}+1 \end{pmatrix}$$
models. The elements of this set are formed by deleting $N^{(t)}+1$ variables from model $\boldsymbol{\gamma}$. The condition $p_\gamma \geq N^{(t)}+1$ must hold to form this neighbourhood set.

- $\boldsymbol{\gamma}^{\mathbf{0}}$: This is a set containing neighbouring models of dimension p_γ and includes
$$|\boldsymbol{\gamma}^{\mathbf{0}}| = \begin{pmatrix} p_\gamma \\ N^{(t)}+1 \end{pmatrix} \times \begin{pmatrix} p - p_\gamma \\ N^{(t)}+1 \end{pmatrix}$$

models. The elements of this set are formed by swapping $2\times(N^{(t)}+1)$ variables of the vector $\boldsymbol{\gamma}$. The conditions $p-p_\gamma \geq N^{(t)}+1$ and $p_\gamma \geq N^{(t)}+1$ must hold to form this neighbourhood set.

We choose uniformly one of the three moves if $p_\gamma \geq N^{(t)}+1$ (otherwise the addition move is chosen) and then draw the proposed model $\boldsymbol{\gamma'}$ uniformly from the corresponding set. The model proposal for the efficient constructed jump proposal algorithms omits the last neighbourhood set $\gamma^{\mathbf{0}}$ since they consider moves from $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma'}$ such that the dimension of $\boldsymbol{\Theta}_\gamma$ is different from the dimension of $\boldsymbol{\Theta}_{\gamma'}$.

The choice of N and π can either be pre-specified or be tuned using short pilot MCMC runs. The parameter π determines the proportion of local to global moves. Small value of π yields more local moves and large value of π more global ones. In the case of $\pi = 0$, the model proposal reduces to the local model proposal which extends the Brown *et al* (1998b) one and randomly chooses to either add or delete a single explanatory variable or to swap two explanatory variables. The corresponding three distinct neighbourhood sets in this case are those used in the shotgun stochastic search algorithm of Hans *et al* (2007).

4 Simulation Results

We apply the transdimensional MCMC samplers described in Section 3 to four datasets from DNA microarray expression studies. Table 1 shows the name of the dataset, the sample size, the number of gene expression variables and each disease group sample size for each dataset. The Arthritis dataset consists of rheumatoid

Dataset	n	p	1st Group	2nd Group
Arthritis	31	755	7	24
Colon Tumour	62	1224	40	22
Leukemia	72	3571	25	47
Prostate	136	10150	59	77

Table 1: Sample size, number of gene expression variables and disease group sample size for each dataset

arthritis and osteoarthritis groups. The Colon Tumour dataset contains tumour and normal colon groups. The Leukemia dataset consists of samples from patients with either acute lymphoblastic leukemia or acute myeloid leukemia and finally the Prostate dataset has prostate tumour and nontumour groups. Detailed descriptions

of the experiments and analysis of those datasets can be found respectively in Sha *et al* (2003), Alon *et al* (1999), Armstrong *et al* (2002) and Singh *et al* (2002).

We set $h = 100$, which leads to a normal prior on the intercept α that is centred at 0 and has a large variance. The gene expression levels have been pre-processed and have a similar scale across the datasets thus it is reasonable to use the same value of c. We choose $c = 5$ which is the value chosen by Sha *et al* (2004) using their guideline method that employs the total relative precision of prior to posterior. We use mean prior model size equal to 5 since models with few genes are expected to give good discrimination.

Table 2 shows the acceptance rate for the Holmes and Held algorithm (Section 3.1.1) with the MC^3 proposal for each dataset using a 500 000 iteration run. This shows the low between-model acceptance rates of the MC^3 algorithm in variable selection problems with large p. The acceptance rate also decreases with the number

Dataset	$\tilde{A}$
Arthritis	1%
Colon Tumour	0.6%
Leukemia	0.2%
Prostate	0.06%

Table 2: The MC^3 acceptance rate for some gene expression datasets

of gene expression variables. This clearly indicates that the MC^3 algorithm is not an efficient algorithm for these problems. Similarly any MCMC algorithm with model proposal that uses the symmetric kernel (3.15) as the dimension-changing move will also not be efficient.

Each MCMC sampler described in Section 3 was run with five different parameters settings of the general model proposal mentioned in Section 3.2. The parameter settings were $\pi = 0, 0.25, 0.5, 0.75, 0.95$ and $N = 4$ in each case. When $\pi = 0$ we randomly choose to either add or delete a single variable or swap two variables and this is the local model proposal. As π increases, we will increase the number of variables we propose to add, delete or swap on average. The maximum number of variables to add or delete is 4 and the maximum number of variables to swap is 8. All the MCMC samplers were run for 500 000 iterations and the first 100 000 draws were discarded to form the burn-in period. Furthermore we thinned the MCMC samplers by steps of 5 to reduce the dependence between the simulated values. The programs were written in Matlab 7.0.1 and run on a desktop PC.

The posterior gene inclusion probabilities are estimated by the ergodic average

$$\hat{\pi}(\gamma_j = 1|\boldsymbol{y}) = \frac{1}{T}\sum_{i=1}^{T}\gamma_j^{(i)}, \qquad j = 1, \ldots, p \tag{4.16}$$

and these estimates for the four datasets are shown in Figure 1. All algorithms give quite similar estimates. In all cases we find a few genes that have significantly higher inclusion probabilities than the others. The highest gene inclusion probability increases with the sample size and number of variables when the posterior distribution will become concentrated on fewer models. Furthermore many variables are highly correlated when p is increasing and therefore the inclusion probability will be spread among many competing models that includes the "best" gene (i.e the gene with the highest posterior inclusion probability). Thus, many more models containing the "best" gene have posterior probability far from 0. Consequently, only a few genes with high gene inclusion probability are distinguished and all the others have almost zero inclusion probabilities when p increases. For example, only one gene of the Prostate dataset and two of the Leukemia dataset have posterior inclusion probability over 0.1 and all the others have probability near zero.

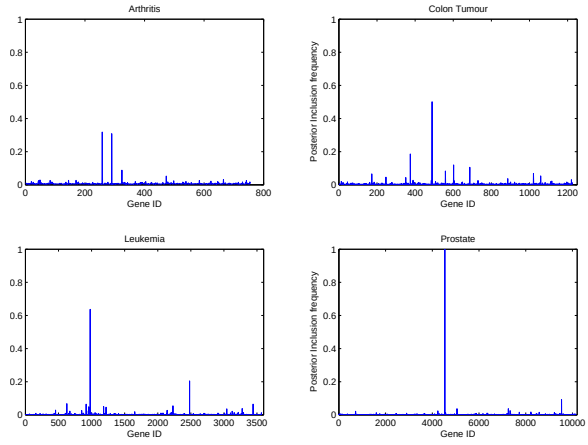

Figure 1: Estimated posterior gene inclusion probabilities for some datasets. We have used the H-H algorithm with a local model proposal

We want to compare the efficiency of the following MCMC algorithms:

1. H-H : Holmes and Held algorithm (Section 3.1.1)
2. AG-LA : Automatic generic sampler with Laplace approximation (Section 3.1.2)

3. AG-IWLS : Automatic generic sampler with Iterated Weighted Least Squares approximation (Section 3.1.2)

4. Z-O : Zeroth Order (Section 3.1.3)

5. F-O : First Order (Section 3.1.3)

6. S-O : Second Order (Section 3.1.3)

7. C-M : Conditional Maximisation (Section 3.1.3)

The efficiency of these algorithms can be compared by monitoring the MCMC output for various parameters. We used the auxiliary variable $\boldsymbol{z}$ since it is continuous and its value is largely determined by the choice of model (the data will only indicate the sign of each element of $\boldsymbol{z}$). A large value of the integrated autocorrelation time τ_i for the i-th component z_i of $\boldsymbol{z}$ is an indication that the MCMC algorithm is not mixing well with respect to the model space. An estimate of τ_i for each z_i was computed using both the initial positive and initial monotone sequence estimators defined by Geyer (1992). We calculated the mean m of τ_i's for each chain and estimated the effective sample size by $\text{ESS} = \frac{T}{m}$ where T is the MCMC sample size after the burn-in and thinning (in this case, T=80 000). A Monte Carlo estimate calculated using a chain with effective sample size k will have the same variance as one calculated using an independent sample of size k.

Table 3 presents the between-model acceptance rate, the effective sample size, the CPU time in seconds and the relative efficiency over the H-H algorithm for each MCMC algorithm with a local model proposal (i.e $\pi = 0$). The last column of Table 3 records the relative efficiency of the MCMC algorithms over the H-H one having standardized for CPU run time. This is defined by

$$\text{R.E} = \frac{\text{ESS(sampler)}}{\text{CPU(sampler)}} \Big/ \frac{\text{ESS(H-H)}}{\text{CPU(H-H)}}.$$

The H-H algorithm always has the lowest acceptance rate. Furthermore some algorithms have high between-model acceptance rate, for example the F-O, S-O and C-M algorithms achieve 66% for the Arthritis dataset and 59% for the Leukemia dataset. The higher order and conditional maximization methods have higher acceptance rates because they do not consider any swap move. However the acceptance rate seems to decrease with the sample size of the dataset because the posterior model distribution becomes less flat. Furthermore when n is large we have a lot

Arthritis

Method	$\tilde{A}$	ESS	CPU	R.E
H-H	41%	5298	3907	1
AG-LA	57%	8602	15936	0.4
AG-IWLS	38%	8421	4574	**1.4**
Z-O	59%	6349	3544	**1.3**
F-O	67%	6400	12949	0.4
S-O	66%	6452	4481	**1.1**
C-M	66%	6723	13089	0.4

Colon Tumour

Method	$\tilde{A}$	ESS	CPU	R.E
H-H	36%	8421	4919	1
AG-LA	49%	10127	16483	0.4
AG-IWLS	36%	8889	5654	0.9
Z-O	41%	9091	3846	**1.4**
F-O	50%	9412	13045	0.4
S-O	50%	9091	4924	**1.1**
C-M	50%	9412	13459	0.4

Leukemia

Method	$\tilde{A}$	ESS	CPU	R.E
H-H	28%	2759	5122	1
AG-LA	46%	3587	18221	0.4
AG-IWLS	38%	3944	6820	**1.1**
Z-O	52%	2963	4881	**1.1**
F-O	59%	3125	15317	0.4
S-O	59%	2996	6060	0.9
C-M	59%	3226	15371	0.4

Prostate

Method	$\tilde{A}$	ESS	CPU	R.E
H-H	25%	7207	8013	1
AG-LA	43%	11268	26907	0.5
AG-IWLS	37%	10000	8636	**1.4**
Z-O	29%	7921	5610	**1.6**
F-O	36%	7692	18802	0.5
S-O	36%	7692	6548	**1.3**
C-M	36%	7767	16982	0.5

Table 3: The acceptance rate, the effective sample size, the CPU time in seconds and the Relative Efficiency over the H-H algorithm for each MCMC sampler with a local model proposal

of information about the regression coefficients and the imputed variable $\boldsymbol{z}$ of the H-H algorithm may not be well-supported under the proposed model $\boldsymbol{\gamma}'$. For these two reasons the H-H algorithm may yield a low acceptance rate for and lead to an inefficient exploration of the model space. Therefore we recommend not using the H-H algorithm with local model proposal when the sample size of the data set is large because this may result in a low acceptance rate.

The Automatic Generic samplers have the highest effective sample sizes followed by the efficient jump proposals and the H-H algorithm. However the AG-LA sampler is computationally expensive. The most efficient method (taking into account the CPU times) are AG-IWLS, Z-O and S-O, followed by the H-H algorithm. The effect becomes more marked as n increases. Therefore we suggest using one of these three samplers when the dataset sample size is large.

Table 3 also shows that the posterior model distribution of the Prostate dataset is less flat than the Arthritis one since the acceptance probabilities of the Prostate dataset are smaller. Therefore the Prostate dataset larger n provides a lot of information about the models. On the other hand the Prostate dataset larger number of highly correlated variables is expected to spread this information among many competing models. The result (a more pronounced posterior distribution) suggests that n is more influential than p. Table 3 also indicates that the Colon Tumour and

Leukemia datasets have quite similar acceptance probabilities and posterior model distributions even though 2300 more variables are included in the Leukemia dataset. Therefore n influences the posterior model distribution, whereas the influence of p is less clear.

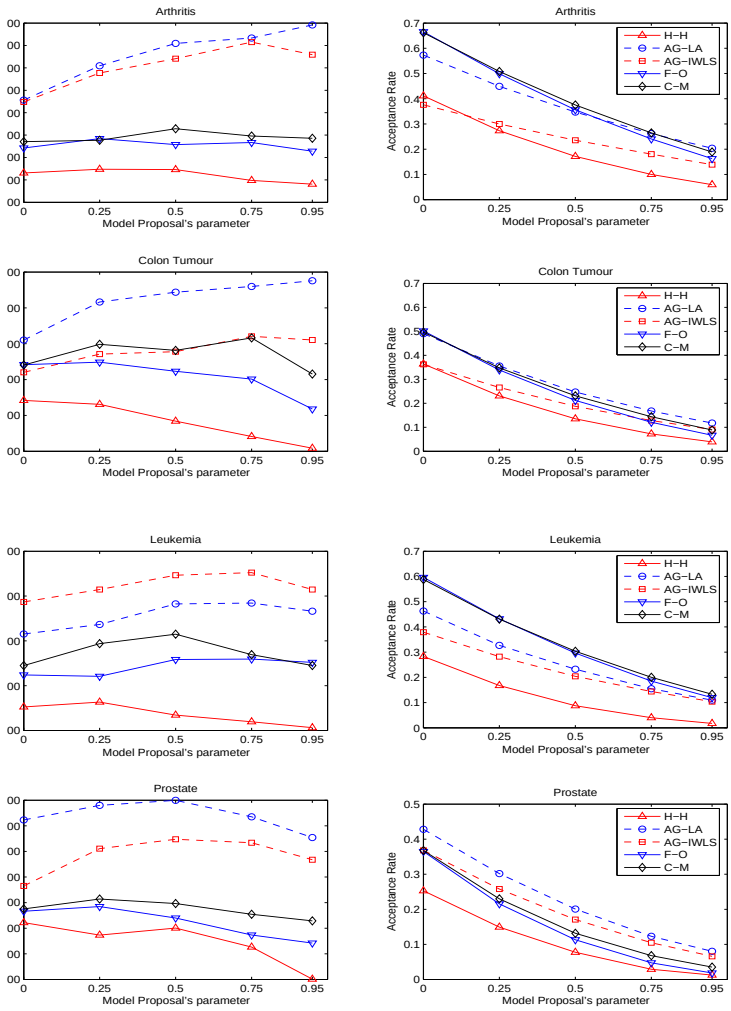

Figure 2: Effective sample size and acceptance rate of the MCMC methods for some data sets. We have used five different model proposal parameters: H-H (solid upper triangle), AG-LA (dashed circle), AG-IWLS (dashed square), F-O (solid down triangle) and C-M (solid diamond)

We now consider using the more general model proposal distributions introduced in Section 3.2. Figure 2 shows how the general model proposal improves the ESS

of the algorithms (left-hand panels), even though it decreases the between-model acceptance rate (right-hand panels). The local proposal (when $\pi = 0$) rarely gives the highest ESS (the exception is the H-H algorithm). More specifically the ESS of the AG-LA sampler is increasing with π for the Arthritis and Colon Tumour datasets. Both the AG-LA and AG-IWLS samplers have maximum ESS if $\pi = 0.75$ or $\pi = 0.5$ with the Leukemia and Prostate datasets respectively. The C-M method gets an optimum ESS if $\pi = 0.5$ when it is applied to the Leukemia dataset. Furthermore, in the Prostate dataset where the acceptance rate for all algorithms is below 10% for $\pi = 0.95$ the F-O and C-M samplers have an optimum ESS if $\pi = 0.25$. It is interesting to note that the optimum ESS is obtained when acceptance rates are between 15% and 25%, which is consistent with standard theory for Metropolis-Hastings random walk proposals (see *e.g.* Roberts and Rosenthal 2001).

The transdimensional MCMC algorithms can be ordered according to their ESS from Figure 2. The Automatic Generic samplers have the highest ESS, followed by the efficiently constructed jump proposals and the H-H algorithm. The AG-IWLS sampler has only slightly lower ESS than the AG-LA sampler even though it uses only rough estimates of the first and second moments of the posterior distribution. Note that the H-H algorithm stands out as having the smallest ESS in combination with the smallest acceptance rate.

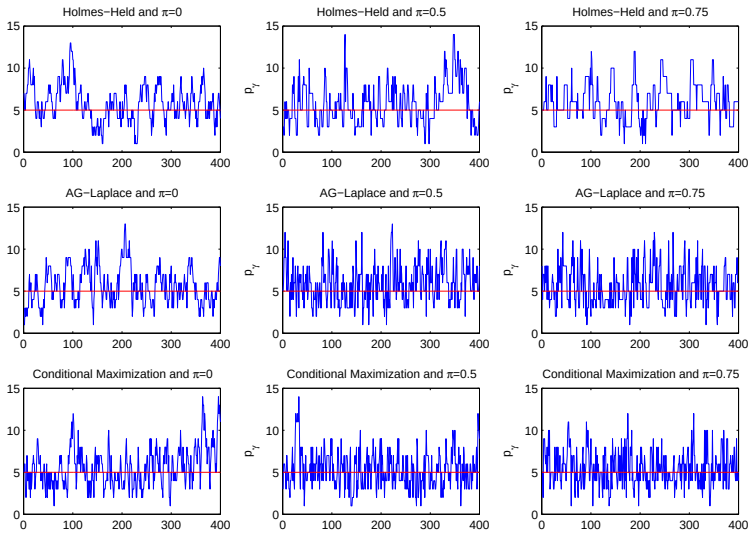

Figure 3: Trace plots of model size for some algorithms with different model proposal's parameters. The Arthritis dataset has been used

The improvements of the general model proposal in the efficiency of the algorithms are also illustrated in Figure 3. It displays the trace plots of the model sizes for the first 400 models sampled after the burn-in period for some algorithms. It is easy to see an improvement in the mixing of the AG-LA sampler if $\pi = 0.5$ or $\pi = 0.75$. There are also improvements in the mixing of both the H-H algorithm with $\pi = 0.5$ and the C-M algorithm with $\pi = 0.5$ and $\pi = 0.75$.

We suggest using the general model proposal with all algorithms except the H-H algorithm, as it will lead to better exploration of the model space and an increase in the ESS. The increase is more pronounced when n is small and the acceptance rate for local model proposals is high. Our applications suggest that the optimum ESS is obtained when the model proposal parameters are chosen to give an acceptance rates between 15% and 25%, which can be achieved by careful tuning of π. The results for the ESS computed on the basis of the intercept α are also quite similar.

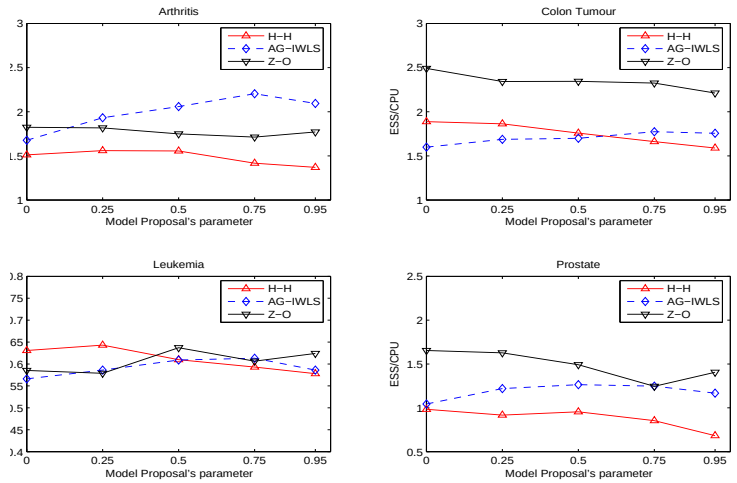

Figure 4: The ESS standardized for the CPU run time using five different model proposal's parameters: H-H (solid upper triangular), AG-IWLS (solid diamond) and Z-O (solid down triangular)

Figure 4 displays the effective sample size standardized by the CPU run time of the AG-IWLS, Z-O and H-H samplers. We only show the results of the most efficient sampler from each group of methods. The AG-IWLS and Z-O samplers are more efficient than the H-H sampler when n is small in the Arthritis dataset (shown in the top left-hand panel). More specifically the AG-IWLS improves the efficiency by 45% for $\pi = 0.5$ and by 55% for $\pi = 0.75$. When the sample size

is moderate as it is for the Colon Tumour and Leukemia datasets, the AG-IWLS and H-H samplers have similar efficiency. The Z-O is more efficient for one of these two datasets. When the sample size is large, as it is for the Prostate dataset, the AG-IWLS and Z-O samplers are at least 30% more efficient than the H-H sampler. Therefore we suggest using the AG-IWLS sampler when the sample size is small and the Z-O sampler when the sample size is large.

The Sha *et al* (2004) and H-H algorithms have the same between-model acceptance rates, however the former is computationally less efficient since sampling from an $n-$variate truncated normal needs more computational time than from n univariate truncated normals and a $p_\gamma-$variate normal (with p_γ typically much smaller than n). Therefore we have omitted the Sha *et al* (2004) algorithm from the comparison study.

5 Discussion

In this paper we have applied existing transdimensional MCMC algorithms to Bayesian variable selection for probit models with $p >> n$, which jointly update the model and the auxiliary variables. The first is the Automatic Generic sampler described by Green (2003). We have compared the Laplace approximation to the first and second moments of the regression coefficient's posterior distribution to rougher estimates from the modified Iterative Weighted Least Square algorithm (Gamerman 1997). The latter sampler has similar mixing to the one using the Laplace approximation but has much lower computational cost. The other transdimensional MCMC algorithms are the higher order and conditional maximization methods introduced by Brooks *et al* (2003). All these algorithms avoid conditioning on auxiliary variables in the model update and tend to mix better than the algorithm of Holmes and Held (2006), which jointly updates the model and the model parameters.

We have also developed a general model proposal that splits the addition-deletion move and combines local moves with more global ones by changing a block of variables simultaneously. The proposal can be "tuned" by the expected number of variables to be changed. This proposal leads to higher effective sample size than the local model proposal for all the transdimensional samplers except the Holmes-Held algorithm. The optimum effective sample size is obtained when acceptance rates are tuned to fall in the range 15% to 25%, which can be achieved by tuning a parameter of the proposal. The development of methods analogous to Adaptive Markov

Chain, see *e.g.* Atchadé and Rosenthal (2005), to tune this parameter would be an interesting direction for future research.

We find that the Automatic Generic samplers have the highest effective sample size followed by the efficiently constructed jump proposals and the Holmes-Held algorithm. If we take computing time into account the Automatic Generic sampler using Iterative Weighted Least Squares is most efficient for small sample sizes ($n \leq 40$) and the Zeroth Order sampler of Brooks *et al* (2003) is most efficient for large sample sizes ($n \geq 120$).

References

Albert, J. and S. Chib (1993): "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88, 669-679.

Alon, U., N. Barkai, and D. A. Notterman (1999): "Broad patterns of gene expression revealed by clustering analysis of tumour and normal colon tissues probe by oligonucleotide array," *Proceedings of the National Academy of Sciences of the United States of America*, 96, 6745-6750.

Armstrong, S. A., J. E. Staunton and L. B. Silverman (2002): "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, 30, 41-47.

Atchadé, Y. F. and J. S. Rosenthal (2005): "On adaptive Markov chain Monte Carlo algorithms," *Bernoulli*, 5, 815-828.

Brooks, S. P., P. Giudici and G. O. Roberts (2003): "Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions," *Journal of the Royal Statistical Society B*, 65, 3-55.

Brown, P. J., M. Vanucci and T.Fearn (1998a): "Multivariate Bayesian variable selection and prediction," *Journal of the Royal Statistical Society B*, 60, 627-641.

Brown, P. J., M. Vanucci and T.Fearn (1998b): "Bayesian wavelength selection in multicomponent analysis," *Journal of Chemometrics*, 12, 173-182.

Chipman, H., E. I. George and R. E. McCullogh (2001): "The practical implementation of Bayesian model selection," in *Model Selection*, ed. P.Lahiri, Hayward, CA:IMS, 67-134.

Denison, D. G. T., C. C. Holmes, B. K. Mallick and A. F. M. Smith (2002): *Bayesian Methods for Nonlinear Classification and Regression*, Chichester: John Wiley & Sons.

Dudoit, S., J. Fridlyand and T. P. Speed (2002): "Comparison of discrimination methods for the classification of tumous using gene expression data," *Journal of the American Statistical Association*, 97. 77-87.

Fernández, C., E. Ley and M. F. J. Steel (2001): "Benchmark priors for Bayesian model averaging," *Journal of Econometrics*, 100, 381-427.

Gamerman, D. (1997): "Sampling from the posterior distribution in generalized linear mixed models," *Statistics and Computing*, 7, 57-68.

Geweke, J. (1991): "Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities," *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, 571-578. Alexandria, Virginia: American Statistical Association.

Geyer, C. J. (1992): "Practical Markov chain Monte Carlo," *Statistical Science*, 7, 473-511.

Golub, T. R., D. K. Slonim, P. Tamayo and C. Huard (1999): "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 531-537.

Green, P. J. (1995): "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711-732.

Green, P. J. (2003): "Trans-dimensional Markov chain Monte Carlo," in *Highly Structured Stochastic Systems*, eds. Green, P.J, N.L. Hjord and S.Richardson, Oxford, U.K.: Oxford University Press, 179-198.

Hans, C., A. Dobra and M. West (2007): "Shotgun stochastic search for "large p" regression," *Journal of the American Statistical Association*, 102, 507-516.

Holmes, C. C. and L. Held (2006): "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian Analysis*, 1, 145-168.

Lee, K. E., N. Sha, R. Dougherty, M. Vannucci and B. K. Mallick (2003): "Gene selection: A Bayesian variable selection approach," *Bioinformatics*, 19, 90-97.

Ley, E. and M.F.J. Steel (2007): "On the effect of prior assumptions in Bayesian Model Averaging with applications to growth regression," *Journal of Applied Econometrics*, forthcoming .

Madigan, D. and J. York (1995): "Bayesian graphical models for discrete data," *International Statistical Review*, 63, 215-232.

Mitchell, T. J. and J. J. Beauchamp (1988): "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023-1032.

Nguyen, D. V. and D. M. Rocke (2002): "Tumor classification by partial least squares using microaary gene expression data," *Bioinformatics*, 18, 39-50.

Raftery, A. E, D. Madigan and J. A. Hoeting (1997): "Bayesian model averaging for linear regression models," *Journal of the American Statistical Assocation*, 92, 179-191.

Roberts, G. O. and J. S. Rosenthal (2001): "Optimal scaling of various Metropolis-Hastings algorithms," *Statistical Science*, 16, 351-367.

Sha, N., M. Vanucci, P. J. Brown, M. Trower and G. Amphlett (2003): "Gene selection in arthritis classification with large-scale microarray expression profiles," *Comparative and Functional Genomics*, 4, 171-181.

Sha, N., M. Vanucci, M. G. Tadesse, P. J. Brown, I. Dragoni, N. Davies, T. C. Roberts, A. Contestabile, M. Salmon, C. Buckley and F. Falciani (2004): "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage," *Biometrics*, 60, 812-819.

Singh, D., P. G. Febbo and K. Ross (2002): "Gene expression correlates of clinical prostate cancer behaviour," *Cancer cell*, 1, 203-209.

Sisson, S. (2005): "Transdimensional Markov chains: A decade of progress and future perspectives," *Journal of the American Statistical Association*, 100, 1077-1089.

Yeung, K. Y., R. E. Bumgarner and A. E. Raftery (2005): "Bayesian model averaging: development of an improved multi-class gene selection and classification tool for microarray data," *Bioinformatics*, 21, 2394-2402.