Transdimensional Sampling Algorithms for Bayesian Variable Selection in Classification Problems with Many More Variables Than Observations

Demetris Lamnisos, Jim E. Griffin and Mark F. J. Steel*

Abstract

Model search in probit regression is often conducted by simultaneously exploring the model and parameter space, using a reversible jump MCMC sampler. Standard samplers often have low model acceptance probabilities when there are many more regressors than observations. Implementing recent suggestions in the literature leads to much higher acceptance rates. However, high acceptance rates are often associated with poor mixing of chains. Thus, we design a more general model proposal that allows us to propose models "further" from our current model. This proposal can be tuned to achieve a suitable acceptance rate for good mixing. 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). We also implement other previously proposed samplers and compare the efficiency of all methods on some gene expression data sets. Finally, the results of these applications lead us to propose guidelines for choosing between samplers.

Key Words: Data augmentation, Gene expression data, Probit model, Reversible jump sampler, Transdimensional Markov chain.

*Demetris Lamnisos is PhD Student, Department of Statistics, University of Warwick, Coventry, U.K., CV4 7AL (Email: D.S.Lamnisos@warwick.ac.uk). Jim E. Griffin is Lecturer, Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, U.K., CT2 7NF (Email: J.E.Griffin-28@kent.ac.uk) and Mark F.J. Steel is Professor, Department of Statistics, University of Warwick, Coventry, U.K. (Email: M.F.Steel@stats.warwick.ac.uk).

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 is to identify this small subset of genes that are linked to the molecular mechanism underlying the disease. Typically, this is complicated by the very large number of potential subsets and the high correlation between many expression levels.

Most variable selection methods in the literature are univariate in the sense that each candidate gene is considered individually (see *e.g.* Dudoit *et al*, 2002). 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, which considers multiple genes simultaneously and, hence, naturally accounts for dependence between genes. However, the standard Bayesian approach to model selection, described by *e.g.* Chipman *et al* (2001), encounters two related problems when applied to the probit model with many explanatory variables. Firstly, the marginal likelihood for each model is not available in analytic form and, secondly, the number of candidate models is very large, prohibiting the exhaustive calculation of the posterior model distribution.

There are at least two different approaches that address these problems. The first approach efficiently identifies a reduced set of good models and uses an approximation to compute the marginal likelihood for each 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 approximated by the Laplace method. The second approach applies Markov chain Monte Carlo (MCMC) 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).

In this paper, we will develop and implement 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. 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 implementing existing forms of reversible jump samplers that jointly update the model and the auxiliary variables. 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 Squares method described by Gamerman (1997). The other algorithms that we apply to this setting 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 contribution 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. Since a Metropolis random walk with local proposals and high acceptance rate is often associated with poor mixing, we generalize this model proposal by adding, deleting or swapping blocks of several variables. These more global moves lead to lower acceptance rates but should lead to 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. The code and data are freely available at `http://www.amstat.org/publications/jcgs`.

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 $\pi(y_i = 1)$ is modelled by $y_i|\eta_i \sim \text{Bernoulli}(\Phi(\eta_i))$ where Φ is the cumulative distribution function of a standard normal random variable and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)'$ is a vector of linear predictors modelled as $\boldsymbol{\eta} = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ an $n \times p$ matrix whose (i, j)-th entry is the measurement of the j-th

covariate for the i-th individual, $\mathbf{1}$ 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 *e.g.* 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 $\boldsymbol{\gamma}$ containing p_γ variables is defined by $\boldsymbol{\eta} = \alpha\mathbf{1} + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma$, 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 $\boldsymbol{\gamma}$ which usually has the 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 $c\mathbf{I}_{p_\gamma}$ (where $\mathbf{I}_q$ is the identity matrix of order q), which is the ridge prior. 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. For 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 is regarded as a common parameter for 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 $\pi(\gamma) = w^{p_\gamma}(1-w)^{p-p_\gamma}$ and p_γ is binomially distributed as $\text{Bin}(p, w)$. Therefore the model size has prior mean pw and variance $pw(1-w)$.

This Bayesian approach to variable selection for the probit model accounts for dependence between explanatory variables and simpler models are favored over more complex ones when

comparable fits are provided. 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 since w plays the main role in inducing a size penalty and c induces regulation on the regression coefficients.

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.

We are not interested in selecting any particular model, but will conduct inference on quantities of interest (such as the gene inclusion probabilities and predictive distributions) on the basis of inference averaged over all models with the posterior model probabilities, *i.e.* we conduct Bayesian model averaging (as in *e.g.* Sha *et al*, 2004).

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 absolutely 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 $\boldsymbol{z} = (z_1, \ldots, z_n)'$ are introduced and

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases} \\ \boldsymbol{z} &\sim N_n(\tilde{\mathbf{X}}_\gamma \boldsymbol{\theta}_\gamma, \mathbf{I}_n), \end{aligned} \tag{3.2}$$

so that 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 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.

Alternatively, a Gibbs sampler for $\boldsymbol{z}, \boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}$ can be used. 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 $\pi(\boldsymbol{\gamma}|\boldsymbol{z})\, \pi(\boldsymbol{\theta}_\gamma|\boldsymbol{z}, \boldsymbol{\gamma})$. 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 standard update steps for $\boldsymbol{z}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}, \boldsymbol{y}$ and $\boldsymbol{\theta}_\gamma|\boldsymbol{z}, \boldsymbol{\gamma}$:

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 $(\tilde{\mathbf{X}}_\gamma \boldsymbol{\theta}_\gamma)_i$ and variance 1 truncated to $(0, \infty)$ if $y_i = 1$ or $(-\infty, 0)$ otherwise.

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 $\boldsymbol{\theta}_\gamma|\boldsymbol{z}, \boldsymbol{\gamma} \sim N_{p_\gamma+1}\left((\tilde{\mathbf{X}}_\gamma'\tilde{\mathbf{X}}_\gamma + \mathbf{H}_\gamma^{-1})^{-1}\tilde{\mathbf{X}}_\gamma'\boldsymbol{z}, (\tilde{\mathbf{X}}_\gamma'\tilde{\mathbf{X}}_\gamma + \mathbf{H}_\gamma^{-1})^{-1}\right)$, where $\mathbf{H}_\gamma$ is a diagonal matrix with h as the first element and c as the p_γ remaining ones. 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 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 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 $\min\{1, A[(\boldsymbol{\theta}_{\gamma}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]\}$. Here

$$A[(\boldsymbol{\theta}_{\gamma}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})] = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})\, \pi(\boldsymbol{\theta}_{\gamma'}|\boldsymbol{\gamma'})\, \pi(\boldsymbol{\gamma'})\, q_{\gamma'}(\boldsymbol{u}_{\gamma'})\, q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})}{\pi(\boldsymbol{y}|\boldsymbol{\theta}_{\gamma}, \boldsymbol{\gamma})\, \pi(\boldsymbol{\theta}_{\gamma}|\boldsymbol{\gamma})\, \pi(\boldsymbol{\gamma})\, q_{\gamma}(\boldsymbol{u}_{\gamma})\, q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})} \left| \frac{\partial g(\boldsymbol{\theta}_{\gamma}, \boldsymbol{u}_{\gamma})}{\partial(\boldsymbol{\theta}_{\gamma}, \boldsymbol{u}_{\gamma})} \right|, \quad (3.3)$$

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 of Green's algorithm is as follows:
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 $\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma'}) = \min\{1, A[(\boldsymbol{\theta}_{\gamma}^{(t)}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]\}$, where $A[(\boldsymbol{\theta}_{\gamma}^{(t)}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]$ is given by (3.3). Otherwise $\boldsymbol{\theta}_{\gamma}^{(t+1)} = \boldsymbol{\theta}_{\gamma}^{(t)}$.

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 (3.3) reduces to

$$A[(\boldsymbol{\theta}_{\gamma}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})] = \frac{\pi(\boldsymbol{\gamma'})\, q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})\, \pi(\boldsymbol{z}|\boldsymbol{\gamma'})}{\pi(\boldsymbol{\gamma})\, q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})\, \pi(\boldsymbol{z}|\boldsymbol{\gamma})} \quad (3.4)$$

The above acceptance probability is independent of both current and proposed parameter states and it is similar to the acceptance probability of a 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 the Holmes and Held algorithm is:
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 $\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma'}) = \min\{1, A[\boldsymbol{\gamma} \to \boldsymbol{\gamma'}]\}$, where $A[\boldsymbol{\gamma} \to \boldsymbol{\gamma'}]$ is given by (3.4).

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 is seen from (3.2), 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 $\boldsymbol{\theta}_\gamma = \boldsymbol{\mu}_\gamma + \boldsymbol{B}_\gamma \boldsymbol{\nu}$ with $\boldsymbol{\mu}_\gamma$ approximating the mean of $\pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{y})$ and $\boldsymbol{B}_\gamma$ approximating the Cholesky decomposition of the covariance matrix of $\pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{y})$. Proposing a new model $\boldsymbol{\gamma'}$, we set a new vector $\boldsymbol{\theta}_{\gamma'}$ to be:

$$\boldsymbol{\theta}_{\gamma'} = \begin{cases} \boldsymbol{\mu}_{\gamma'} + \boldsymbol{B}_{\gamma'}(\boldsymbol{R}\boldsymbol{\nu})_1^{p_{\gamma'}+1} & \text{if} \quad p_{\gamma'} < p_\gamma \\ \boldsymbol{\mu}_{\gamma'} + \boldsymbol{B}_{\gamma'}\boldsymbol{R}\boldsymbol{\nu} & \text{if} \quad p_{\gamma'} = p_\gamma \\ \boldsymbol{\mu}_{\gamma'} + \boldsymbol{B}_{\gamma'}\boldsymbol{R}\begin{bmatrix} \boldsymbol{\nu} \\ \boldsymbol{u}_\gamma \end{bmatrix} & \text{if} \quad p_{\gamma'} > p_\gamma. \end{cases} \tag{3.5}$$

Here $(\cdot)_1^m$ denotes the first m component of a vector, $\boldsymbol{R}$ is a fixed orthogonal matrix of order $\max\{p_\gamma + 1, p_{\gamma'} + 1\}$ 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. 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 $\min\{1, A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \to (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]\}$ and (3.3) takes the form

$$A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})] = \frac{\pi(\boldsymbol{\gamma'}, \boldsymbol{\theta}_{\gamma'}|\boldsymbol{y})\ q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})\ |\boldsymbol{B}_{\gamma'}|}{\pi(\boldsymbol{\gamma}, \boldsymbol{\theta}_\gamma|\boldsymbol{y})\ q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})\ |\boldsymbol{B}_\gamma|} \times \begin{cases} q_{\gamma'}(\boldsymbol{u}_{\gamma'}) & \text{if} \quad p_{\gamma'} < p_\gamma \\ 1 & \text{if} \quad p_{\gamma'} = p_\gamma \\ q_\gamma(\boldsymbol{u}_\gamma)^{-1} & \text{if} \quad p_{\gamma'} > p_\gamma \end{cases} \tag{3.6}$$

Since $\boldsymbol{R}$ is orthogonal it does not play any role in this calculation. In the case $p_{\gamma'} < p_\gamma$, $\boldsymbol{u}_{\gamma'}$ is derived in two steps. First, $\boldsymbol{\theta}_{\gamma'}$ is found using the first line of (3.5) and then $\boldsymbol{u}_{\gamma'}$ follows from the third line of (3.5) by interchanging $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$.

The motivation is that high transition probabilities may be achieved when $\pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{y})$ is 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 is:
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 (3.5).
4. Jump to the model $\boldsymbol{\gamma'}$ and set $\boldsymbol{\theta}_{\gamma'}^{(t+1)} = \boldsymbol{\theta}_{\gamma'}$ with probability $\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma'}) = \min\{1, A[(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]\}$, where $A[(\boldsymbol{\theta}_\gamma^{(t)}, \boldsymbol{\gamma}) \rightarrow (\boldsymbol{\theta}_{\gamma'}, \boldsymbol{\gamma'})]$ is given by (3.6). Otherwise $\boldsymbol{\theta}_\gamma^{(t+1)} = \boldsymbol{\theta}_\gamma^{(t)}$.

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 Squares (IWLS) 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 inverse of the negative Hessian matrix at $\hat{\boldsymbol{\mu}}_\gamma$, respectively. This method solves an optimization problem in each iteration and is thus not computationally 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 IWLS algorithm to find rough estimates of the first and second moments. This algorithm finds the posterior mode $\hat{\boldsymbol{\mu}}_\gamma$ by iterating

$$\boldsymbol{\mu}_\gamma^{(k)} = \left(\mathbf{H}_\gamma^{-1} + \tilde{\mathbf{X}}_\gamma' \boldsymbol{W}\left(\boldsymbol{\mu}_\gamma^{(k-1)}\right)\tilde{\mathbf{X}}_\gamma\right)^{-1} \tilde{\mathbf{X}}_\gamma' \boldsymbol{W}\left(\boldsymbol{\mu}_\gamma^{(k-1)}\right)\tilde{\boldsymbol{y}}\left(\boldsymbol{\mu}_\gamma^{(k-1)}\right)$$

until convergence, where $\mathbf{H}_\gamma$ is the prior covariance matrix of $\boldsymbol{\theta}_\gamma$, $\tilde{\boldsymbol{y}}\left(\boldsymbol{\mu}_\gamma^{(k-1)}\right)$ is a vector of transformed observations and $\boldsymbol{W}\left(\boldsymbol{\mu}_\gamma^{(k-1)}\right)$ is a diagonal matrix of weights. The inverse 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 probit model the vector of transformed observations $\tilde{\boldsymbol{y}}\left(\boldsymbol{\mu}_\gamma^{(k-1)}\right)$ is defined as

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

and the diagonal matrix of weights $\boldsymbol{W}\left(\boldsymbol{\mu}_\gamma^{(k-1)}\right)$ 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 $\eta_i = \tilde{\mathbf{x}}_{\gamma i}\boldsymbol{\mu}_\gamma^{(k-1)}$, $\tilde{\mathbf{x}}_{\gamma i}$ is the ith row of the design matrix $\tilde{\mathbf{X}}_\gamma$ and ϕ is the probability density function of the standard normal. We can use one or more iteration cycles of this modified IWLS method to find rough estimates of the first and second moments of $\pi(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{y})$. In our implementation described in `http://www.amstat.org/publications/jcgs` we use a single iteration from a starting value based on a least squares regression using a rough estimate of $\boldsymbol{z}$. This method proves computationally more efficient than the Laplace approximation.

3.1.3 Efficient Construction of Reversible Jump Proposal Densities

Brooks *et al* (2003) discuss 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 Taylor series expansion of the acceptance probability (3.3) 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 $g(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma(\boldsymbol{v}))$, where $\boldsymbol{u}_\gamma$ is a general proposal transformation of some canonical random $\boldsymbol{v}$. 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 $\boldsymbol{v}$. 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 $\boldsymbol{u}_\gamma(\boldsymbol{v}) = \boldsymbol{\mu}+\sigma\boldsymbol{v}$ and $\boldsymbol{v} \sim N_{p_{\gamma'}-p_\gamma}(\mathbf{0}, \mathbf{I}_{p_{\gamma'}-p_\gamma})$. The between-model map is set to the identity, *i.e.* $g(\boldsymbol{\theta}_\gamma, \boldsymbol{u}_\gamma(\boldsymbol{v})) = (\boldsymbol{\theta}_\gamma, \boldsymbol{\mu} + \sigma\boldsymbol{v})$. 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 $\big(p_{\gamma'} + 1\big)$-dimensional model with parameters $(\boldsymbol{\theta}_\gamma, \mathbf{0})$. Thus the likelihood drops out of equation (3.3). Furthermore, the Jacobian in (3.3) is

$$\left|\frac{\partial(\boldsymbol{\theta}_\gamma, \boldsymbol{\mu} + \sigma\boldsymbol{v})}{\partial(\boldsymbol{\theta}_\gamma, \boldsymbol{v})}\right| = \sigma^{p_{\gamma'}-p_\gamma}. \tag{3.7}$$

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 (3.3) around the centering point $c(\boldsymbol{\theta}_\gamma)$.

Automatic Proposals

These methods automatically specify the parameters of a proposal using the form of the acceptance probability. The parameters are 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 (3.3) equals 1, that is

$$A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \to (c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})] = 1. \tag{3.8}$$

and the first r derivatives of the logarithm of the acceptance probability are required to be equal to the zero vector at $c(\boldsymbol{\theta}_\gamma)$, that is

$$\nabla^k \log A[(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) \to (c(\boldsymbol{\theta}_\gamma), \boldsymbol{\gamma'})] = \mathbf{0}, \qquad k = 1, \ldots, r$$

Here the partial derivatives are taken with respect to $\boldsymbol{v}$. 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 methods are named after the chosen value of r. The zeroth-

order method specifies the scale of the proposal transformation $\boldsymbol{u}_\gamma(\boldsymbol{\upsilon}) = \sigma\boldsymbol{\upsilon}$, while $\boldsymbol{\mu}$ is set to $\mathbf{0}$. Given that $c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, \mathbf{0}), \boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma)) = \mathbf{0}$, and $\boldsymbol{b}(\boldsymbol{\theta}_\gamma) = \mathbf{0}$ and using (3.7), (3.8) leads to

$$\sigma = \left(\frac{c^{(p_{\gamma'} - p_\gamma)/2} \pi(\boldsymbol{\gamma}) q(\boldsymbol{\gamma'}|\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma'}) q(\boldsymbol{\gamma}|\boldsymbol{\gamma'})} \right)^{\frac{1}{p_{\gamma'} - p_\gamma}} \tag{3.9}$$

where c is the hyperparameter that determines the prior covariance matrix $\mathbf{V}_\gamma = c\mathbf{I}_{p_\gamma}$ in (2.1).

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. The first order method imposes an $p_{\gamma'} - p_\gamma + 1$ dimensional constraint on the proposal. The location and scale of the proposal transformation $\boldsymbol{u}_\gamma(\boldsymbol{\upsilon}) = \boldsymbol{\mu} + \sigma\boldsymbol{\upsilon}$ are the solutions to the above system of equations. Given that $c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, \mathbf{0}), \boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma)) = \mathbf{0}$ and $\boldsymbol{b}(\boldsymbol{\theta}_\gamma) = -\boldsymbol{\mu}/\sigma$, the system of equations is written as

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

where $\mathbf{X}_{p_{\gamma'} - p_\gamma}$ is a $(p_{\gamma'} - p_\gamma) \times n$ matrix with as 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 $c(\boldsymbol{\theta}_\gamma)$, but no longer imposes (3.8). Now, there could be more constraints than the proposal parameters needed to be determined. Also, this method is computationally demanding when $p_{\gamma'} > p_\gamma + 1$ since the constraint on the Hessian matrix considerably increases the number of equations to be solved. If $p_{\gamma'} = p_\gamma + 1$, the second-order method involves two constraints and only the two parameters μ and σ need to be determined. Using the fact that $c(\boldsymbol{\theta}_\gamma) = (\boldsymbol{\theta}_\gamma, 0), u_\gamma(b(\boldsymbol{\theta}_\gamma)) = 0$, and $b(\boldsymbol{\theta}_\gamma) = -\mu/\sigma$, the solution to the

implied system of equations 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 the $1\times n$-dimensional 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. In our implementation of the second-order method, we will only consider adding up to two variables.

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

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

where $\mathbf{X}_u$ is a $(p_{\gamma'}-p_\gamma)\times n$ matrix with the new explanatory variables proposed to be included. We use that $c(\boldsymbol{\theta}_\gamma)=(\boldsymbol{\theta}_\gamma,\boldsymbol{\mu}),\boldsymbol{u}_\gamma(\boldsymbol{b}(\boldsymbol{\theta}_\gamma))=\boldsymbol{\mu}$, and $\boldsymbol{b}(\boldsymbol{\theta}_\gamma)=\mathbf{0}$ to derive that the scale of the proposal is given by:

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

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

$$q(\boldsymbol{\gamma}'|\boldsymbol{\gamma}) = \frac{1}{p} \quad \text{if} \quad \sum_{i=1}^{p} |\gamma_i' - \gamma_i| = 1 \tag{3.10}$$

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. 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 spends 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. Thus, a large number of adding moves are rejected yielding a low between-model acceptance rate.

Brown *et al* (1998b) extended the model proposal (3.10). 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 (3.10) and therefore yields similar low between models acceptance rates.

We introduce a new model proposal in two stages. Firstly, we split the MC^3 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 candidate model $\boldsymbol{\gamma}'$ differs from the current one $\boldsymbol{\gamma}$ by either one or two variables (with a swap move). This local model proposal will often yield high between-model acceptance rates when applied to problems with $p >> n$. One possible reason is that $\boldsymbol{\gamma}'$ will be similar to $\boldsymbol{\gamma}$ in terms of model fit when p is large and many explanatory variables are either redundant or highly correlated. A second reason is that for small sample size n the posterior distribution will be relatively flat and a large number of models are well supported by the data.

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 a block of variables simultaneously. Thus, it is designed to enable the fast exploration of the model space. First, we 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, *i.e.* $N^{(t)} \sim \text{Bin}(N-1, \pi)$ 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.

- $\boldsymbol{\gamma^0}$: This is a set containing neighbouring models of dimension p_γ and includes

$$|\boldsymbol{\gamma^0}| = \begin{pmatrix} p_\gamma \\ N^{(t)}+1 \end{pmatrix} \times \begin{pmatrix} p-p_\gamma \\ N^{(t)}+1 \end{pmatrix}$$

  models. The elements of this set are formed by swapping $(N^{(t)}+1)$ variables of the vector $\boldsymbol{\gamma}$ for the same number of previously excluded variables. The conditions $p-p_\gamma \geq N^{(t)}+1$ and $p_\gamma \geq N^{(t)}+1$ must both hold to form $\boldsymbol{\gamma^0}$.

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 efficiently 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 values of π yield more local moves and large values 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 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).

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

We use a normal prior on the intercept α that is centred at 0 and has a large variance ($h = 100$). 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. The value of c regulates the amount of shrinkage and does affect gene inclusion probabilities in this context, but values that are relatively close (*e.g.* $c = 10$) lead to very similar conclusions on the efficiency of the algorithms.

Running the Holmes and Held algorithm (Section 3.1.1) with the MC^3 proposal for each dataset for 500,000 iterations reveals the low between-model acceptance rates of the MC^3 algorithm in variable selection problems with large p. The acceptance rates range from 1% for the Arthritis data to 0.06% for the Prostate data and decrease with the number of gene expression variables. This clearly indicates that the MC^3 algorithm is not an efficient algorithm for these problems. Generally, any MCMC algorithm with the symmetric model proposal (3.10) as the dimension-changing move is inefficient.

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 after the first 100,000 draws were discarded and the chains were thinned by only recording every 5th draw. The samplers were implemented in Matlab 7.0.1 and run on a desktop PC.

The posterior gene inclusion probabilities are estimated by the ergodic average

$$\hat{\pi}(\gamma_j = 1|\boldsymbol{y}) = \frac{1}{T}\sum_{i=1}^{T}\gamma_j^{(i)}, \qquad j = 1, \ldots, p \tag{4.11}$$

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. As sample size and the number of variables increase, the posterior inclusion probabilities become more concentrated on a smaller number of genes.

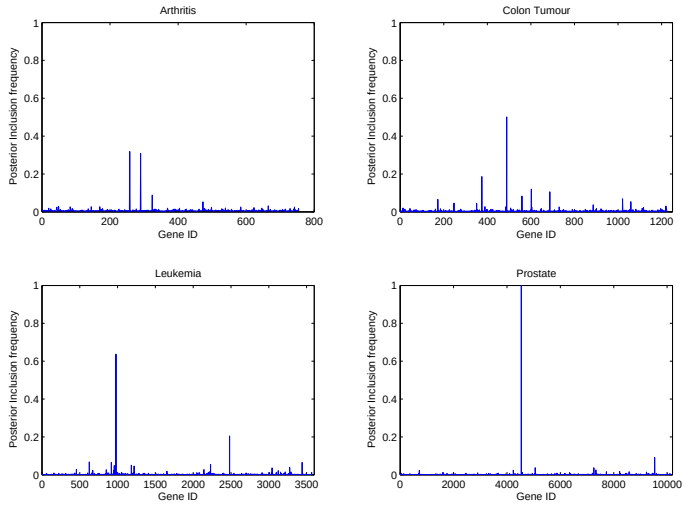

Figure 1: Estimated posterior gene inclusion probabilities for some datasets. We have used the H-H algorithm with a local model proposal

We compare the efficiency of the following MCMC algorithms:

H-H : Holmes and Held algorithm (Section 3.1.1)

AG-LA : Automatic generic sampler with Laplace approximation (Section 3.1.2)

AG-IWLS : Automatic generic sampler with IWLS approximation (Section 3.1.2)

Z-O : Zeroth-Order (Section 3.1.3)

F-O : First-Order (Section 3.1.3)

S-O : Second-Order (Section 3.1.3)

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 components γ_i of $\boldsymbol{\gamma}$ since the posterior gene inclusion

probabilities (4.11) are the main quantities of interest in gene expression studies. Results using the auxiliary variables z_i instead lead to very similar conclusions. An estimate of the integrated autocorrelation time τ_i for each γ_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 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.

Arthritis

Method	$\tilde{A}$	ESS	CPU	R.E
H-H	41%	8498	1449	1
AG-LA	57%	13363	4552	0.5
AG-IWLS	52%	11784	1771	**1.1**
Z-O	53%	9679	1361	**1.2**
F-O	66%	13100	4728	0.5
S-O	66%	12819	1621	**1.3**
C-M	66%	13165	4235	0.5

Colon Tumour

Method	$\tilde{A}$	ESS	CPU	R.E
H-H	36%	7950	2017	1
AG-LA	49%	11402	5103	0.6
AG-IWLS	47%	10537	2284	**1.2**
Z-O	36%	6467	1629	**1**
F-O	50%	9698	4886	0.5
S-O	49%	9391	1973	**1.2**
C-M	50%	9646	4843	0.5

Leukemia

Method	$\tilde{A}$	ESS	CPU	R.E
H-H	28%	6190	2901	1
AG-LA	46%	10705	10067	0.5
AG-IWLS	36%	7808	3229	**1.1**
Z-O	47%	9200	2165	**2**
F-O	60%	11874	5854	0.9
S-O	59%	11865	2590	**2**
C-M	59%	12063	5678	1

Prostate

Method	$\tilde{A}$	ESS	CPU	R.E
H-H	26%	8012	2987	1
AG-LA	43%	13125	7438	0.7
AG-IWLS	36%	11385	2988	**1.4**
Z-O	24%	6766	2145	**1.2**
F-O	37%	9423	6927	0.5
S-O	36%	9328	2583	**1.3**
C-M	35%	9441	6220	0.6

Table 2: The acceptance rate $\tilde{A}$, 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

Table 2 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 2 records the relative efficiency of the MCMC algorithms over the H-H one having standardized for CPU run time:

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

The H-H algorithm almost always has the lowest acceptance rate. Some algorithms have high between-model acceptance rate, *e.g.* 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. 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 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 $\boldsymbol{\gamma}'$. Thus, the H-H algorithm may result in a lower acceptance rate than the Automatic Generic Samplers and may lead to an inefficient exploration of the model space.

The Automatic Generic samplers tend to have the highest effective sample sizes followed by the efficient jump proposals and the H-H algorithm. However, the AG-LA, F-O and C-M samplers are computationally expensive. The most efficient methods (taking into account the CPU times) are AG-IWLS, Z-O and S-O, followed by the H-H algorithm. Therefore, we suggest using the AG-IWLS, Z-O or S-O samplers if we stick with a local model proposal.

Table 2 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 larger number of highly correlated variables of the Prostate dataset is expected to spread this information among more competing models. The result (a more pronounced posterior distribution) suggests that n is more influential than p. Table 2 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.

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 most 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 AG-LA, AG-IWLS and C-M samplers have maximum ESS for $\pi = 0.5$ and $\pi = 0.25$ with the Arthritis and Colon Tumour datasets, respectively. The F-O sampler gets an optimum ESS for $\pi = 0.25$ when it is applied to Arthritis dataset. Furthermore, in the Leukemia dataset all the samplers except the H-H one have maximum ESS if $\pi = 0.25$. Finally, the AG-LA and AG-IWLS have an optimum ESS if $\pi = 0.25$ when they are applied to the Prostate dataset. It is interesting to note that the optimum ESS is obtained when acceptance rates are between 25% and 40%, which is consistent with standard

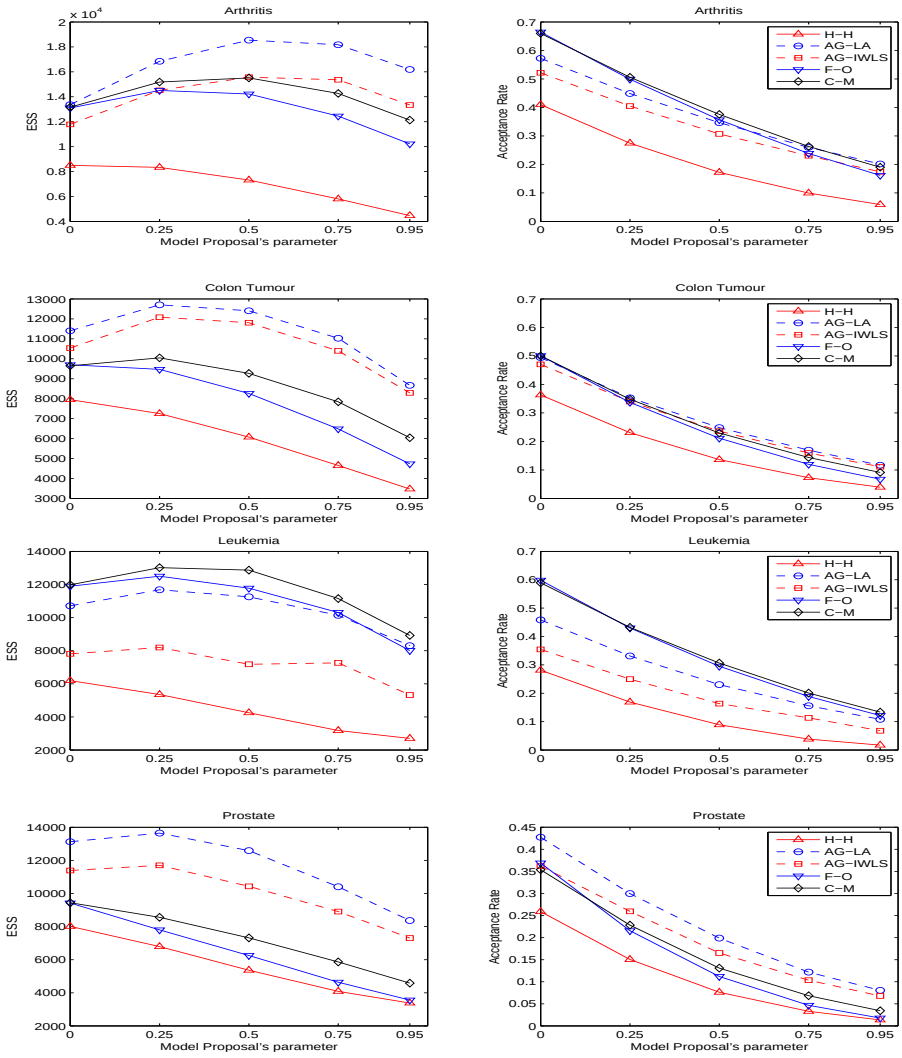

Figure 2: Effective sample size and acceptance rate of the MCMC methods for some data sets. We have used five different model proposal parameters: H-H (solid with triangle), AG-LA (dashed with circle), AG-IWLS (dashed w. square), F-O (solid w. inverted triangle) and C-M (solid w. diamond)

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 tend to 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 throughout.

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 25% and 40%, which can be achieved by careful tuning of π. The results for the ESS computed on the basis of the auxiliary variable $\boldsymbol{z}$ and the intercept α are also quite similar.

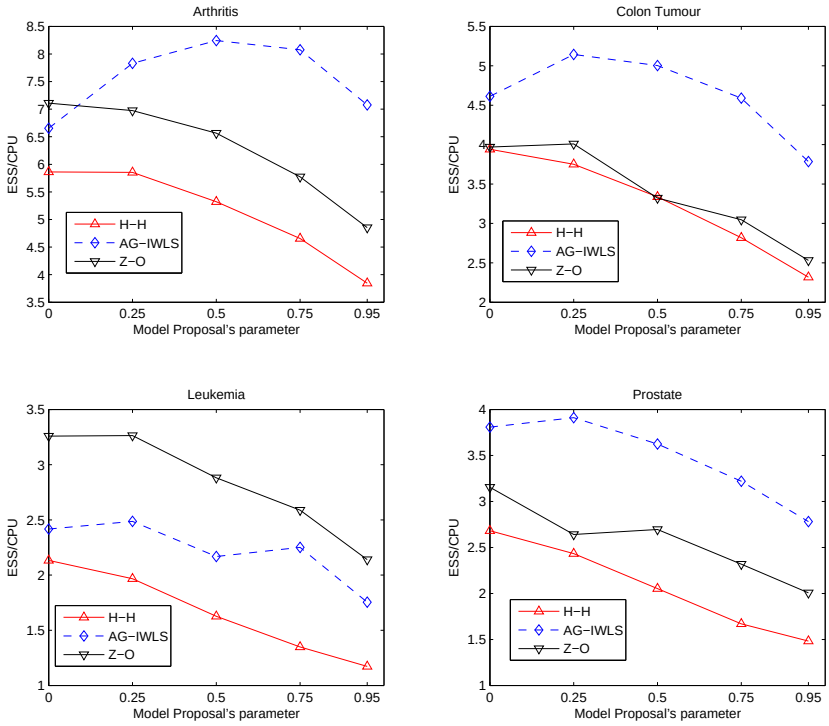

Figure 3: The ESS standardized for the CPU run time using five different model proposal parameters: H-H (solid with triangle), AG-IWLS (solid w. diamond) and Z-O (solid w. inverted triangle)

Figure 3 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 in all the datasets. More specifically, the AG-IWLS improves the efficiency by 55% for $\pi = 0.5$ and by 75% for $\pi = 0.75$ in the Arthritis dataset. Furthermore, the AG-IWLS

improves the efficiency by 50% for $\pi = 0.5$ and by 65% for $\pi = 0.75$ in the Colon Tumour dataset. The Z-O and AG-IWLS samplers are at least 50% more efficient than the H-H sampler in the Leukemia and Prostate datasets respectively. Therefore, we suggest using the AG-IWLS and Z-O samplers.

The Sha *et al* (2004) and H-H algorithms have the same between-model acceptance rates, but the former is computationally less efficient since sampling from an $n-$variate truncated normal needs more computational effort 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 posterior moments of the regression coefficients with rougher estimates from the modified Iterative Weighted Least Squares 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 fall in the range 25% to 40%, 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 and the Zeroth-Order sampler of Brooks *et al* (2003) are the most efficient.

Acknowledgements: we are grateful to the Editor, an Associate Editor and two referees for constructive comments.

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 U.S.A.*, 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. McCulloch (2001): "The practical implementation of Bayesian model selection," in *Model Selection*, ed. P.Lahiri, Hayward, CA:IMS, 67-134.

Dudoit, S., J. Fridlyand and T. P. Speed (2002): "Comparison of discrimination methods for the classification of tumours 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.

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

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.

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.

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.