Bayesian Analysis (2009) 4, Number 1, pp. 85–118

A Review of Bayesian Variable Selection Methods: What, How and Which

R.B. O'Hara* and M. J. Sillanpää†

Abstract. The selection of variables in regression problems has occupied the minds of many statisticians. Several Bayesian variable selection methods have been developed, and we concentrate on the following methods: Kuo & Mallick, Gibbs Variable Selection (GVS), Stochastic Search Variable Selection (SSVS), adaptive shrinkage with Jeffreys' prior or a Laplacian prior, and reversible jump MCMC. We review these methods, in the context of their different properties. We then implement the methods in BUGS, using both real and simulated data as examples, and investigate how the different methods perform in practice. Our results suggest that SSVS, reversible jump MCMC and adaptive shrinkage methods can all work well, but the choice of which method is better will depend on the priors that are used, and also on how they are implemented.

Keywords: Variable Selection, MCMC, BUGS

1 Introduction

An important problem in statistical analysis is the choice of an optimal model from a set of *a priori* plausible models. In many analyses, this reduces to the choice of which subset of variables should be included into the model. This problem has exercised the minds of many statisticians, leading to a variety of algorithms for searching the model space and selection criteria for choosing between competing models (e.g. Miller 2002; Broman and Speed 2002; Sillanpää and Corander 2002). In the Bayesian framework, the model selection problem is transformed to the form of parameter estimation: rather than searching for the single optimal model, a Bayesian will attempt to estimate the posterior probability of all models within the considered class of models (or in practice, of all models with non-negligible probability). In many cases, this question is asked in variable-specific form: i.e. the task is to estimate the marginal posterior probability that a variable should be in the model.

At present, the computational method most commonly used for fitting Bayesian models is Markov chain Monte Carlo (MCMC) technique (Robert and Casella 2004). Variable selection methods are therefore needed that can be implemented easily in the MCMC framework. In particular, having these models implemented in the BUGS language (either in WinBUGS or OpenBUGS) means that the methods can easily be slotted into different models. The purpose of this paper is to review the different methods that

*Department of Mathematics and Statistics, University of Helsinki, Finland, mailto:bob.ohara@helsinki.fi

*Department of Mathematics and Statistics, University of Helsink i, Finland, http://www.rni.helsinki.fi/~mjs/

 DOI:10.1214/09-BA403

have been suggested for variable selection, and to present BUGS code for their implementation, which may also help to clarify the ideas presented. Some of these methods have been reviewed by Dellaportas et al. (2000). We do not consider some methods, such as Bayesian approximative computational approaches (e.g. Ball 2001; Sen and Churchill 2001) or methods based on calculating model choice criteria, such as DIC (Spiegelhalter et al. 2002), because these are only feasible to use with a maximum of dozens of candidate models. We also omit the machine learning literature focusing on finding *maximum a posteriori* estimates for parameters (e.g. Tipping 2004).

Although methods for variable selection are reviewed here, this should not be taken to imply that they should be used uncritically. In many studies, the variables in the regression have been chosen because there is a clear expectation that they will influence the response, and the problem is one of inferring the strength of influence. For these studies, the best strategy may therefore be to fit the full model, and then interpret the sizes of the posterior estimates of the parameters in terms of their importance. In other studies the purpose is more exploratory: seeing what the analysis throws out from a large number of candidates. This is not always an unreasonable approach. One clear example where this is a sensible way to proceed is in gene mapping, where it is assumed that there are only a small number of genes that have a large effect on a trait, while most of the genes have little or no effect. The underlying biology is therefore sparse: only a few factors (i.e. genes) are expected to influence the trait. The distinction between this case, and many other regression problems is in the prior distribution of effect sizes: for gene mapping, the distribution is extremely leptokurtic, with a few large effects, but most regression coefficients being effectively zero. The prior is also exchangeable over the loci (i.e. we are ignorant *a priori* of where any influential gene might be located). Conversely, in many studies the expectation may be of a slow tapering of effects, with no clear tail to the distribution, and often more substantial information about the likely size of the effects. This implies that a different set of priors should be considered in these two situations: this paper examines different options for the exploratory case where the prior may be leptokurtic.

This review is structured so that we first set out the general regression model. We then describe the different variable selection methods, and some of their properties. Then we describe three examples, using simulated and real data sets, to illustrate the performance of the different methods. Finally, we wrap up by discussing the relative merits of these methods, and indicate when different methods might be preferred. BUGS code for the methods is given in Supplementary Material online, as are some supplementary figures.

2 The Bayesian Variable Selection Methods

2.1 Description of sparse selection problem

The problem is the familiar regression problem of trying to explain a response variable with a (large) number of explanatory variables (whether continuous or discrete factors). The aim is to select a small subset of the variables whilst controlling the rate of false detection, so that it can be inferred that these variables explain the a large fraction of the variation in the response. We may have some *a priori* knowledge or expectation that only a small proportion of candidates are truly affecting the outcome, and ideally this information should be taken into account in the variable selection. The optimal degree of sparseness and how many false detections are allowed is very problem-specific.

One aspect of the problem is the well known trade-off between bias and variance. In general, a large set of variables is desirable for prediction and a small set of variables (that have a meaningful interpretation) for inference. Another aspect that influences the number of variables in the model is the number of observations in the data set. As a rule of thumb for shrinkage methods, one can safely consider only problems where there are maximally 10-15 times more candidates than observations (Zhang and Xu 2005; Hoti and Sillanpää 2006). However, where the true and safe upper limit exists, is very problem specific and depends on degree of correlation (co-linearity) among the candidate variables.

2.2 Regression model

To keep the presentation simple, assume that the task is to explain an outcome y_i for individual i $(i = 1, ..., N)$ using p covariates with values $x_{i,j}, j = 1, .., p$. Naturally, these may be continuous or discrete (dummy) variables. Given a vector of regression parameters $\theta = (\theta_j)$ of size p, the response y_i is modeled as a linear combination of the explanatory variables $x_{i,j}$:

$$y_i = \alpha + \sum_{j=1}^{p} \theta_j x_{i,j} + e_i. \tag{1}$$

Here, α is the intercept and $e_i \sim N(0, \sigma^2)$ are the errors. We can assume more generally that y_i is a member of the exponential family of distributions, giving us a generalized linear model. In such a case, we take the usual link function $g(\cdot)$ so that $E(g(y_i)) = \eta_i$ and η_i equals the right hand side of the linear model (1) without the error term. The variable selection procedure can be seen as one of deciding which of the regression parameters, the θ_js, are equal to zero. Each θ_j should therefore have a "slab and spike" prior (Miller 2002), with a spike (the probability mass) either exactly at or around zero, and a flat slab elsewhere. For this, we may use an auxiliary indicator variable I_j (where $I_j = 1$ indicates presence, and $I_j = 0$ absence of covariate j

in the model) to denote whether the variable is in the slab or spike part of the prior. A second auxiliary variable, the effect size β_j, is also needed for most of the methods, where $\beta_j = \theta_j$ when $I_j = 1$ (e.g. by defining $\theta_j = I_j\beta_j$). When $I_j = 0$, the variable β_j can be defined in several ways, and this leads to the differences between the methods outlined below. For the moment we will assume $\theta_j = I_j\beta_j$, to simplify the explanation, and change it later on when needed (e.g. for SSVS below).

Once the model has been set up, it is usually fitted using MCMC, and many of the methods outlined below use a single-site Gibbs sampler to do this. The variable selection part of the model entails estimating I_j and θ_j. From this, the posterior probability that a variable is "in" the model, i.e. the posterior inclusion probability, can simply be calculated as the mean value of the indicator I_j. The methods outlined below vary in how they treat I_j, β_j and θ_j.

2.3 Concepts and Properties

The methodologies of Bayesian variable selection and the differences between them can be best understood using several properties and concepts, which are described here.

Sparseness

The degree of sparseness required, i.e. how complex the model should be is an important property. In some cases, the sparseness may be set according to some optimality criterion (e.g. the best predictive abilities, (Burnham and Anderson 2002)). Taking a decision analytic perspective, we can view the prior as providing a loss function, so unless an objective optimality criterion can be found, it is not clear that one loss function is appropriate for all circumstances. Therefore, some flexibility in the amount of model complexity allowable is needed. An obvious approach to this is to use $P(I_j = 1)$, the prior probability of variable inclusion, to set the sparseness: a smaller value of $P(I_j = 1)$ leading to sparser models. Typically, this will be independent across the js, so that the prior distribution for the number of covariates is binomial, with mean $P(I_j = 1)$. A value of 0.5 for this has been suggested for $P(I_j = 1)$ (e.g. George and McCulloch 1993), which makes all models equiprobable. Whilst this may improve mixing properties of the MCMC sampler and may appear attractive as a null prior, it is informative in that it favours models where about half of the variables are selected. In many cases, only a small proportion of variables are likely to be required in the model, so this prior may often bias the model towards being too complex. The choice of value for $P(I_j = 1)$ is then up to the investigator, in some cases a decision analytic approach may be a good way of eliciting the prior.

Tuning

A practical problem in implementing variable selection is tuning the model (by adjusting different components, such as the prior distribution) to ensure good mixing of the MCMC chains, i.e. letting the sampler jump efficiently between the slab and spike. If single-site updaters are being used, this means updating I_j given a value of β_j. This relies on confounding of I_j and β_j, so that $\theta_j \approx 0$ for both $I_j = 1$ and $I_j = 0$, and hence the updater for I_j can jump between states easily (and θ_j be updated subsequently). In general, this will depend on the prior for θ_j, so the mixing properties of the sampler depend on the prior distributions. This has lead to the suggestion that data-based priors should be designed with the purpose of improving mixing (see references below) or giving good centering and scaling properties (e.g. a fractional prior, Smith and Kohn 2002). Although this is attractive from the computational point of view, it contravenes the Bayesian philosophy, as the prior should be a summary of the beliefs of the analyst (before seing the data), not a reflection of the ability of the fitting algorithms to do their job. One goal of this review is to find out under what circumstances the different methods work efficiently, so that philosophically correct (subjective) priors can be designed properly. This may require a trade-off, with a sub-optimal model being used, in order to accommodate better (philosophically plausible) priors.

Several of the methods below may exhibit problems in the marginal identifiability (i.e. confounding) of variables I_j and β_j. This can occur because almost identical likelihoods can be obtained for $I_j = 1$ and $I_j = 0$ when β_j is near zero, as illustrated above. Deliberate confounding of variables can thus be used as a strategy to improve mixing. Because of this, though, it may be more informative to monitor the posterior of the product $\theta_j = I_j \times \beta_j$ instead of the individual variables (Sillanpää and Bhattacharjee 2005), i.e. monitor the parameter that appears in equation 1.

Global adaptation

A natural Bayesian strategy for building a model would be to place a normal prior on $\beta_j \mid (I_j = 1)$. If the variance of this prior is fixed at a constant, the model for the data would be equivalent to a classical fixed effects model, a terminology we will adopt here. But variable selection approaches can be developed where the variance is estimated as well. In some circumstances, this can be done by extending the model (1) above as a hierarchical model; considering the regression coefficients to be exchangeable and be drawn from a common distribution, e.g. $\beta_j \mid (I_j = 1)$ is drawn from $N(0, \tau^2)$, where τ^2 is an unknown parameter to be estimated. We will follow the terminology in classical statistics and refer to this as a random effect model. This approach has the advantage of helping tuning, for example if we define $\theta_j = \beta_j I_j$, then $\beta_j \mid (I_j = 0)$ will also depend on τ^2. The distribution of θ_j is thus shrunk towards the correct region of the parameter space by the other θ_js. This can help circumvent tuning problems, at the cost of increasing the confounding of I_j and β_j, as is discussed more below.

Local adaptation

Instead of fitting one common parameter for all the regression coefficients, as in global adaptation, one can use different variance parameters for different covariates (or for covariate groups). This helps tuning and allows heterogeneity, i.e., if there are some local characteristics among covariates. An example of this is where $\beta_j \mid (I_j = 1)$ is drawn from $N(0, \tau_j^2)$, where τ_j^2 is an covariate-specific variance parameter to be estimated (cf. Example 2 in (Iswaran and Rao 2005)).

Analytical integration

To speed up the convergence of the MCMC (and mixing with respect to selected covariates) in some of the model selection methods described above, it is possible to analytically integrate over the effects θ and σ^2 in model (1), and then use Gibbs sampling for the indicator variables (George and McCulloch 1997). Fast mixing is possible because updating an indicator does not depend on values of the effect coefficients. If preferred, the posterior for effect coefficients can still be obtained. Additionally in high dimensional problems, the coefficients β_j only need to be updated for covariates with $I_j = 1$ (e.g. Yi 2004).

Bayesian Model Averaging

One characteristic of the Bayesian approach is the ability to marginalize over nuisance parameters. This carries over into model selection, where posterior distributions of parameters (including indicators for models) can be calculated by averaging over all of the other variables, i.e. over the different models. For example, as we will do below, the probability that a single variable should enter a model can be averaged over all of the models. Of course, this is convenient in the MCMC framework as it just means that the calculations can be done on the MCMC output for each indicator (i.e. I_j) separately.

It is also well known that, with many covariates, it is the ones that have a large effect that are selected, even if support for estimated effect size is large by chance (e.g. Miller 2002; Lande and Thompson 1990; Göring et al. 2001). Hence, if the same data is used for estimating both the the model (i.e. variable selection) and individual contributions (effect sizes), overestimation of effect sizes will almost certainly occur. Fortunately, robust estimation of effect sizes can be done in a Bayesian setting by averaging the effect size over several different models (e.g. Ball 2001; Sillanpää and Corander 2002). Given posterior model probabilities, model-specific effect estimates are weighted by the probability of the corresponding model. This is an especially useful technique if there are several competing models which all show high posterior probabilities.

2.4 Variable Selection Methods

The approaches to variable selection can be classified into four categories, with different methods in each category. The structures of the models are summarized in Table 1.

Indicator model selection

The most direct approach to variable selection is to set the slab, $\theta_j \mid (I_j = 1)$ equal to β_j, and the spike, $\theta_j \mid (I_j = 0)$ equal to 0. This approach has spawned two methods, differing in the way they treat $\beta_j \mid (I_j = 0)$:

Kuo & Mallick. The first method simply sets $\theta_j = I_j\beta_j$ (Kuo and Mallick (1998)). This assumes that the indicators and effects are independent *a priori*, so $P(I_j, \beta_j) = P(I_j)P(\beta_j)$, and independent priors are simply placed on each I_j and β_j. The MCMC algorithm to fit the model does not require any tuning, but when $I_j = 0$, the updated value of β_j is sampled from the full conditional distribution, which is its prior distribution. Mixing will be poor if this is too vague, as the sampled values of β_j will only rarely be in the region where θ_j has high posterior support, so the sampler will only rarely flip from $I_j = 0$ to $I_j = 1$. This method has been used for applications in genetics by Uimari and Hoeschele (1997) and with local adaptation by Sillanpää and Bhattacharjee (2005, 2006). Smith and Kohn (2002) use this approach to model sparse covariance matrices for longitudinal data (where they use a Cholesky decomposition of the covariance matrix, which reduces the problem to one of variable selection).

GVS. An alternative model formulation called Gibbs variable selection (GVS) was suggested by Dellaportas et al. (1997), extending a general idea of Carlin and Chib (1995). It attempts to circumvent the problem of sampling β_j from too vague a prior by sampling $\beta_j \mid (I_j = 0)$ from a "pseudo-prior", i.e. a prior distribution which has no effect on the posterior. This is done by setting $\theta_j = I_j\beta_j$ as before, but now the prior distributions of indicator and effect are assumed to depend on each other, i.e. $P(I_j, \beta_j) = P(\beta_j \mid I_j)P(I_j)$. In effect, a mixture prior is assumed for β_j: $P(\beta_j \mid I_j) = (1 - I_j)N(\tilde{\mu}, S) + I_j N(0, \tau^2)$ (here and elsewhere we will loosely use $N(\cdot, \cdot)$ to denote both a normal distribution and its density function), where constants $(\tilde{\mu}, S)$ are user-defined tuning parameters, and τ^2 is a fixed prior variance of β_j. The intuitive idea is to use a prior for $\beta_j \mid (I_j = 0)$ which is concentrated around the posterior density of θ, so that when $I_j = 0$, $P(\beta_j \mid I_j = 1)$ is reasonable large, and hence there is a good probability that the chain will move to $I_j = 1$. The algorithm does require tuning, i.e. $(\tilde{\mu}, S)$ need to be chosen so that good values of β_j are proposed when $I_j = 0$. The data will determine which values are good but without directly influencing the posterior, and hence tuning can be done to improve mixing without changing the model's priors.

Stochastic search variable selection (SSVS)

In this approach, the spike is a narrow distribution concentrated around zero. Here $\theta_j = \beta_j$ and the indicators affect the prior distribution of β_j, i.e., $P(I_j, \beta_j) = P(\beta_j \mid I_j)P(I_j)$. A mixture prior for β is used: $P(\beta_j \mid I_j) = (1 - I_j)N(0, \tau^2) + I_j N(0, g\tau^2)$, where the first density (the spike) is centred around zero and has a small variance. This model gives identifiability for variables I_j and β_j, but in order to obtain convergence the algorithm requires tuning - specification of fixed prior parameters (τ^2 and $g\tau^2$) which are data (or at least context) dependent. Note that unlike in GVS, values of the prior parameters when $I_j = 0$ influence the posterior. Tuning is not easy, as $P(\beta_j \mid I_j = 0)$ needs to be very small but at the same time not too restricted around zero (otherwise Gibbs sampler moves between states $I_j = 0$ and $I_j = 1$ are not possible in practice). The technique was introduced by George and McCulloch (1993) and extended for multivariate case by Brown et al. (1998). It has seen extensive use, for example see Yi et al. (2003); Meuwissen and Goddard (2004) for applications to gene mapping.

Meuwissen and Goddard (2004) introduced (in multivariate context) a random effects variant of SSVS where τ^2 was taken as a parameter to be estimated in the model with own prior, and g was fixed at 100. A natural alternative would be to fix τ^2, and (in effect) estimate g, in practice by placing a prior on the product $g\tau$.

Adaptive shrinkage

A different approach to inducing sparseness is not to use indicators in the model, but instead to specify a prior directly on θ_j that approximates the "slab and spike" shape. Hence, $\theta_j = \beta_j$, with prior $\beta_j \mid \tau_j^2 \sim N(0, \tau_j^2)$, and a suitable prior is placed on τ_j^2 to give the appropriate shape to $P(\beta_j)$. The prior should work by shrinking values of β_j towards zero if there is no evidence in the data for non-zero values (i.e. the likelihood is concentrated around zero). Conversely, there should be practically no shrinkage for data-supported values of covariates that are non-zero. The method is adaptive in the sense that a degree of sparseness is defined by the data, through the way it shrinks the covariates effects towards zero. The degree of sparseness of the model can be adjusted by changing the prior distribution of τ_j^2 (either by changing the form of the distribution or the parameters). Tuning in this way may also affect the mixing of the MCMC chains. A problem is that there is no indicator variable to show when a variable is 'in' the model, however one can be constructed by setting a standardised threshold, c, such that $I_j = 1$ if $\mid \beta_j \mid > c$ (cf. Hoti and Sillanpää 2006).

Jeffreys' prior. A scale invariant Jeffreys' prior, $P(\tau_j^2) \propto \frac{1}{\tau_j^2}$, provides one method for adaptive shrinkage. Theoretically, the resulting posterior is not proper (e.g. Hopert and Casella 1996; ter Braak et al. 2005), although a proper approximation can be made by giving finite limits to $P(\tau_j^2)$ (see below). There is no tuning parameter in the model, which is either good or bad: the slab part of the prior is then uninformative but cannot be adjusted. See Xu (2003) for introduction and application of this method

to gene mapping. See Zhang and Xu (2005) for penalized ML equivalent of the method.

Laplacian shrinkage. An alternative to using the Jeffreys' prior on the variance is to use an exponential prior for τ_j^2 with a parameter μ. After analytical integration over the variance components, we obtain a Laplacian double exponential distribution for $P(\beta_j \mid \mu)$; for details, see Figueiredo (2003). The degree of sparseness is controlled by μ which has a data dependent scale and requires tuning. The random effect variant of the method, where μ is a parameter and has its own prior, is better known as the Bayesian Lasso (Park and Casella 2008; Yi and Xu 2008). The Lasso (Tibshirani 1996) is the frequentist equivalent of this approach.

Model space approach

The models above are developed through placing priors on the individual covariates θ_js. An alternative approach is to view the model as a whole, and place priors on N_v, the number of covariates selected in the model and their coefficients $(\theta_1, \ldots, \theta_{N_v})$, and then allow the choice of which covariate it is that is in the model to be a secondary problem. This approach can reduce to the models above, if the number of covariates in the model is chosen to be binomially distributed with N_{max} equal to the number of candidate covariates, p. However, it is often computationally more convenient to use a lower N_{max}, i.e. to restrict the maximum number of covariates possible. The advantage of this approach is that the likelihood is smaller, as one only needs to sum over the selected variables, replacing the summation in model (1) by $\sum_{k=1}^{N_v} \theta_k x_{i,l_k}$. The number of selected variables, N_v, is then itself a random variable, and sparseness can be controlled through the prior distribution of N_v.

Reversible jump MCMC. Reversible jump MCMC is a flexible technique for model selection introduced by Green (1995), which lets the Markov chain explore spaces of different dimension. For variable selection, the positions (indices) of the selected variables are defined as $l_1, .., l_{N_v}$, and the model is updated by randomly selecting variable j and then proposing either addition to ($N_v := N_v + 1$) or deletion from ($N_v := N_v - 1$) the model of the corresponding effect. The length of the vector of θ_js is therefore not fixed but varies during the estimation. The updating is done using a Metropolis-Hastings algorithm, but with the acceptance ratio adjusted for the change in dimension. The degree of sparseness can be controlled by setting the prior for N_v: using a binomial distribution is then approximately the same as setting a constant, independent, prior for each I_j. Reversible jump MCMC has been applied in many areas, and its use is wider than just selecting variables in a regression. See Sillanpää and Arjas (1998) for application to gene mapping and the paper by Lunn et al. (2006) for a WinBUGS application. See also the brief note of Sillanpää et al. (2004).

Composite model space (CMS). A problem with reversible jump MCMC is that the change in dimension increases the complexity of the algorithm. This can be circumvented by fixing N_{max} to something less than p, but to use indicator variables to allow

Table 1: The qualitative classification of the variable selection methods with respect to speed, mixing and separation in our tested examples (AVE: average, EXC: Excellent).

Method	Link	Prior	Speed	Mixing	Separation
Kuo & Mallick	$\theta_j = I_j\beta_j$	$P(\beta_j)P(I_j)$	SLOW	GOOD	GOOD
GVS	$\theta_j = I_j\beta_j$	$P(\beta_j \mid I_j)P(I_j)$	SLOW	GOOD	GOOD
SSVS	$\theta_j = \beta_j$	$P(\beta_j \mid I_j)P(I_j)$	AVE.	GOOD	GOOD
Laplacian	$\theta_j = \beta_j$	$P(\beta_j \mid \tau_j^2)P(\tau_j^2)$	AVE.	POOR	POOR
Jeffreys'	$\theta_j = \beta_j$	$P(\beta_j \mid \tau_j^2)P(\tau_j^2)$	AVE.	EXC.	GOOD
Reversible Jump	θ	$P(\theta \mid N_v)P(N_v)$	FAST	MIXED	GOOD

covariates to enter or leave the model (with the constraint $\sum_j I_j \leq N_{max}$). As in indicator model selection above, $\theta_j = I_j\beta_j$ and both *a priori* independence or dependence between indicators and effects can be assumed. Because the maximum dimension is fixed, the indicators, I_j, are mutually dependent, with maximum and minimum values of $\sum_j I_j$ being set. Variable selection is then performed by randomly selecting component j and then proposing a change of the indicator value I_j (this corresponds to adding or deleting the component). The prior for the number of components can be set in the same way as in reversible jump MCMC. The method was introduced by Godsill (2001), and has been used by Yi (2004) in a gene mapping application. See Kilpikari and Sillanpää (2003) for a closely related approach from the reversible jump MCMC stand point.

3 Examples of the Methods

The efficiency of using BUGS to fit the different models outlined above was examined by coding each of them for three sets of data: a simulation study and two real data sets, from gene mapping in barley (Tinker et al. 1996) and a classic regression data set of the mortality effects of Pollution (McDonald and Schwing 1973). In following, these three data sets are called Simulated data, Barley data, and Pollution data. The code for the Barley analysis is given in the Appendix, and can easily be modified or extended and used as a part of more complex models.

The different variable selection methods work by specifying the priors for β_j, and possibly other auxiliary variables. Interest lies in both the estimates of the parameters (in particular, whether they are consistent across models) and the behaviour of the MCMC, i.e. how long the runs take, how well the chains mix, etc. For all three examples, short runs were used to estimate running time and quality of mixing, and then a longer run (chosen to give reasonable level of mixing), was used to obtain posterior distributions of the parameters, which could then be examined to see how well the methods classified variables as being included in the models, and also whether the

parameter estimates were similar.

For all three sets of data, the equation (1) formed the basis of the model. However, for the simulated data where a generalized linear model was used, equation 1 gives the expected value on the linear scale for each datum.

3.1 Simulated Data

In any real data set, it is unlikely that the "true" regression coefficients are either zero or large; the sizes are more likely to be tapered towards zero. Hence, the problem is not one of finding the zero coefficients, but of finding those that are small enough to be insignificant, and shrinking them towards zero. This situation was mimicked with simulated data, using a simple Poisson model with over-dispersion. 11 replicated data sets were created, each with a total of 200 individuals i ($i = 1, .., 200$), and 20 covariates with values $x_{i,j}$, $j = 1, .., 20$, were used, and the differences being in the true values of the regression parameters θ_j. The Poisson simulation model is:

$$\eta_i = \alpha + \sum_{j=1}^{p} \theta_j x_{i,j} + e_i, \tag{2}$$

where $log(\lambda_i) = \eta_i$ with the observed counts $y_i \sim Poisson(\lambda_i)$ and the (over-dispersion) errors $e_i \sim N(0, \sigma_e^2)$.

For the simulations, known values of $\alpha = ln(10)$ and $\sigma_e^2 = 0.75^2$ were used. The covariate values, the $x_{i,j}$'s, were simulated independently from a standard normal distribution, N(0,1). The regression parameters, θ_j, were generated according to a tapered model, i.e. $\theta_j = a + b(j/10.5 - 1)$, with $a = 0.3$ and $b = 0, 0.05, 0.1, ..., 0.3$ for each data set: this gave them a mean of 0.3, and a range between 0 and 0.6.

The same model (2), was used to analyse the simulated data sets with prior distributions specified as below and in Table 2.

3.2 Barley Data

The data was taken from the North American Barley Genome Mapping project (Tinker et al. 1996). This was a study of economically important traits in two-row barley (*Hordeum vulgare* L.), using 150 doubled-haploid (DH) lines. We concentrate on phenotypic data on time to heading, averaging over all environments for each line with data from every environment. The marker data, set of discrete covariates $x_{i,j}$, comes from 127 (biallelic) markers covering on seven chromosomes so that two different genotypes are segregating (in equal proportions) at each marker. The model is, in effect, a

127-way ANOVA, with a normally distributed response and 127 two-level factors. Because the model is almost saturated (127 covariates, 150 data points), this is the type of problem where an efficient variable selection scheme is necessary.

Some discrete marker genotypes are missing (in total 5% of the covariates values, with all individuals having at least 79% of their covariate information observed). A model for the missing covariate data (i.e. $x_{i,j}$s) is therefore needed. Because of the design of the crosses, for each covariate, the two alleles (i.e. genotype classes) are equally likely, so we assume that the missing data are missing completely at random (MCAR), and assume $x_{i,j} \sim Bern(0.5)$. For simplicity we assume that the covariates are independent, although in reality dependence will be present as the genetic markers are sometimes close to each other on the chromosome (Fig. 4). A more complex model (e.g. Knapp et al. 1990; Sillanpää and Arjas 1998) would be preferable for a "real" analysis.

3.3 Pollution Data

This is a classic data set for investigating variable selection, and was first presented by McDonald and Schwing (1973). The response variable is the age-adjusted mortality rates in 1963, from 60 metropolitan areas of the US. There are 15 potential predictors, all continuous and here are all standardised to have unit variance. We assume that the errors are normally distributed.

3.4 Priors for all analyses

For each set of data, two sets of priors were used. The first set was chosen to be vague, and the second was chosen to be more informative. In particular, the second set of priors for α and β_j were chosen to be representative of prior knowledge about the range of the effects. A more usual prior for I_j was chosen, so that each model was *a priori* equally likely. The following priors were assumed for all models, the constants used are given in Table 2:

$$\alpha \sim N(0, \sigma_\alpha^2) \tag{3}$$

$$\beta_j \mid (I_j = 1) \sim N(0, \sigma_\beta^2) \tag{4}$$

$$\sigma^2 \sim Inverse - Gamma(10^{-4}, 10^{-4})$$

We did not try local adaptation in any of the methods as it is likely to behave very similarly to adaptive shrinkage. However, we tried two versions of the methods, with and without global adaptation, i.e., varying the way we treated σ_β^2 above. For the fixed

effects model σ_β^2 is given a constant value, but in the random effects model it has a distribution, so the standard deviation is given a uniform prior:

$$\sigma_\beta \sim U(0, 20), \tag{5}$$

where $U(a, b)$ denotes a uniform distribution between a and b (for a justification of this prior, see Gelman (2006)).

The form of the prior distribution for β_j and the indicator I_j depends on which variable selection method is used. Because the adaptive shrinkage method with the Jeffreys' prior has no parameters that can be adjusted to change the degree of sparseness, this model was used as a benchmark for the analysis of each set of data. For this model, the posterior of several of the parameters is bimodal (this corresponds to covariates where $P(I_j = 1 \mid data)$ is not near 0 or 1), and a suitable cut-off, c could be chosen by visual examination, so that $|\beta_j| < c$ would be equivalent to $I_j = 0$ (cf. Hoti and Sillanpää 2006). From this, the number of non-zero components (i.e. number of estimated values of $|\beta_j|$ above c) was estimated and rounded to give a prior for I_j. $P(I_j = 1)(= p)$ and c are also given in Table 2. This approach to prior specification was taken to help give consistency in the comparisons: clearly it should not be used for actual analyses.

3.5 Implementation in BUGS

The models were all implemented in OpenBUGS3.0.2, and run in R through the BRugs package (Thomas et al. 2006). The exception to this was the reversible jump MCMC method, which is not presently available in OpenBUGS, so was run in WinBUGS1.4 through the R2WinBUGS package (Sturtz et al. 2005). The BUGS code for the Barley data analyses is given in the Appendix. A description of the models is given here, values of parameters of the prior distributions are given in Table 2. The following models were run:

No Selection

The model with no model selection was used as a baseline for comparison. The vague priors were essentially those for $I_j = 1, \forall j$, i.e. equation 4 for β for the fixed effect, and equations 4 and 5 for the random effect model (viz., similar to ridge regression).

Kuo & Mallick

The method of Kuo and Mallick (1998) was implemented using I_j as a number (0 or 1), and setting $\theta_j = I_j\beta_j$. A mathematically equivalent implementation would use I_j as an

Table 2: Parameters of prior distributions for different variable selection methods for three different sets of data.

Parameter	Simulated Data		Barley Data		Pollution Data	
	Priors 1	Priors 2	Priors 1	Priors 2	Priors 1	Priors 2
μ_α	0	$log(10)$	0	60	0	950
σ^2_α	10^2	1	10^6	10^2	10^{10}	10^3
σ^2_β	10^2	1	10^6	10^2	10^6	10^3
c	0.07	0.07	0.05	0.05	5	5
p	0.2	0.5	20/127	0.5	0.2	0.5
σ^2_{GVS}	0.25	0.25	4	4	10^2	10^2

indicator:

$$\theta_j = \begin{cases} 0 & \text{if } I_j = 0 \\ \beta_j & \text{if } I_j = 1 \end{cases} \tag{6}$$

This was also investigated, but the performance was the same in either case, so only results from the first implementation are reported. The other priors (e.g. for β_j) are as above, for both the fixed and random effect models.

GVS

For GVS a pseudo-prior is needed for $I_j = 0$, otherwise the model is the same as the Kuo & Mallick model. For this, for both the fixed and random effect models, $\beta_j \mid (I_j = 0) \sim N(0, \sigma^2_{GVS})$ was used.

SSVS

The priors for $\beta_j \mid (I_j = 1)$ are as above for both the fixed and random effect models. Both random effect models suggested above were tried. For the fixed effect model and the first random effect model, for $I_j = 0$ the prior for β_j was constructed so that $P(|\beta_j| < c) < 0.01$, by setting it to be 3 standard deviations away from the mean, i.e. $\beta_j \mid (I_j = 0) \sim N(0, (3 \times c)^2)$. For the second random effect model (i.e. due to Meuwissen and Goddard 2004) we used $g = 10^{-3}$. This second model is referred to as M & G.

Adaptive shrinkage (Jeffreys' prior)

Only a single version of this adaptive shrinkage method is possible. The prior for τ^2_j was $log(\tau^2_j) \sim U(-50, 50)$, for all sets of data, which is an finite approximation to the fully correct method and should cover the realistic range of τ^2_j (approximately 10^{-22} to 10^{22}).

Laplacian shrinkage

A prior on the scale (μ) is needed for this model. For the fixed effect, this was designed so that *a priori* $P(\beta_j > c) \approx p$. This lead to the prior $|\beta_j| \sim Exp(-log(1-p)/c)$. For the random effect version (i.e. the Bayesian LASSO), a uniform prior U(0,20) was placed on μ.

Reversible Jump MCMC

The priors specified above were used for both fixed and random effects. The prior is given on the number of variables, N_v, in the model, so this was a binomial distribution with $P(N_v) \sim Bin(m, p)$.

Composite Model Space

The priors defined above were used, but a maximum of 40 variables was set. The results of the short run for the Barley data showed that composite model space was too slow to be useful, being about three times slower than any of the other methods, and with poor mixing (the fixed effect model had not even converged after 1500 MCMC iterations). Full runs were therefore not attempted.

The speed of the Composite Model Space in BUGS is due to the way that BUGS implements the model, rather than an intrinsic problem with the model. BUGS compiles all logical nodes fully, so that for each variable in the model, the node has to include every covariate in its calculation. Hence, each of the possible combinations of covariates is included, and so the likelihood quickly becomes excessively large. Implementations coded from scratch will therefore be much quicker.

3.6 Comparisons of Methods

The efficiency and mixing properties of the methods were investigated by carrying out short runs. For all of the data, two chains of each model were run for 1000 MCMC iterations after a burn-in of 500 MCMC iterations (except for the random effect variant of Kuo & Mallick model, which required 1000 MCMC iterations to burn-in). The time taken, the effective number of MCMC samples for α (Geyer 1992; Plummer et al. 2008), and the number of runs of 0s and 1s in the chains for each I_j were all recorded. The number of runs is a measure of mixing: more runs indicate better mixing (i.e. more flips between the variable being in the model and not). The fixed effect version of the Kuo & Mallick method with the vague priors was omitted from the comparisons with the simulated data because its performance was not stable.

From the short runs, the full runs were designed to have a burn-in and thinning suf-

ficient to give good mixing (a minimum burn-in of 500 MCMC iterations was used, and thinning to keep between every fifth and every fiftieth MCMC iteration). The choice of thinning depended on the effective number of MCMC samples, the number of runs, and a visual inspection of the MCMC chain histories, to check the mixing. On the basis of the results, the full runs were thinned to every 20 iterations, with the exception of the No Selection (fixed effect version) and adaptive shrinkage with Jeffreys' prior (both thinned to every 10 iterations), and the random effect version of Kuo & Mallick, fixed effect version of SSVS and both reversible jump MCMC methods (all thinned to 50). The same thinning was used for both sets of priors.

The full data were examined to see how well the methods worked. In particular, how efficiently they separated the variables into those to be included in the model, and those to be excluded. Ideally $P(I_j \mid data)$ should be close to 0 or 1, with very few intermediate values. The posterior probabilities can also be represented as Bayes Factors (e.g. Kass and Raftery 1995), and the categories of Kass & Raftery can be used to judge the strength of evidence that a variable should be included in a model. The estimates of the regression parameters should also be consistent across methods: we would expect the different methods to give the same estimates.

Influence of Tuning

As indicated above, some of the methods require tuning to obtain good mixing. These were examined, in particular looking at the following properties by varying the variables:

- the tuning of the pseudo-prior in GVS
- the scale of the "spike" in the fixed effect SSVS
- the scale of c in the M & G model
- the scale of the Laplacian prior

The mixing of the MCMC chain in the GVS and SSVS models was measured by the number of runs in all of the indicators, I_j whilst the effect of the scale of the Laplacian prior was investigated by examining the posterior distribution of the β_js.

4 Results

An overall qualitative assesment of different aspects (computational speed, efficiency of mixing and separation) of the performance of the methods in the three data sets is summarized in Table 1.

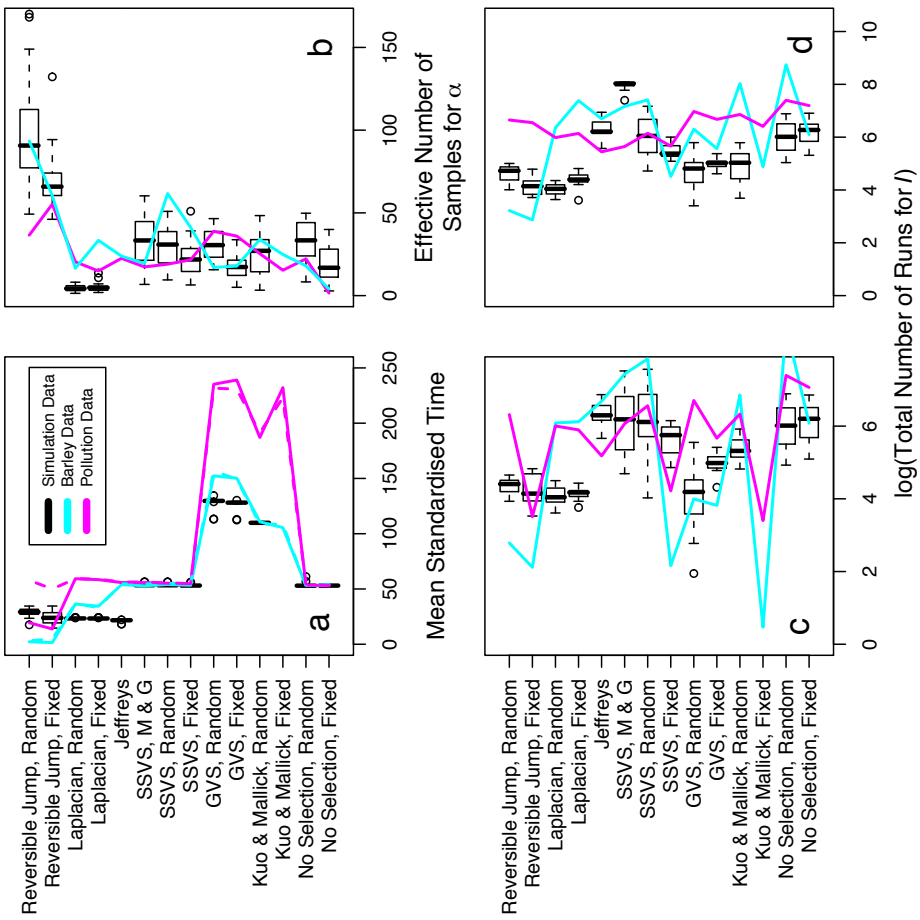

Figure 1: Statistics for short runs for three data sets (simulated data as boxplots, Barley and Pollution data as lines) and 11 variable selection methods. (a) Standardized times to run 1000 MCMC iterations (standardised to have a mean of 1), (b) Estimated effective number of MCMC samples for α (for adaptive shrinkage using Jeffreys' prior applied to Simulated data and all Pollution results > 150), (c) Total numbers of runs for indicator variables, Priors set 1, (d) Total numbers of runs for indicator variables, Priors set 2. The fixed effect variant of Kuo & Mallick method was not run for the simulated data.

4.1 Computational Performance

Summaries of the computational performance of the different methods are plotted in Figure 1. Speedwise, the GVS and Kuo & Mallick methods are slower per iteration than the others, whilst the reversible jump MCMC can be much quicker. The effective number of MCMC samples are all fairly similar, except that the Laplacian shrinkage does less well, and the reversible jump MCMC tends to do better than the other methods. The adaptive shrinkage using Jeffreys' prior performed much better than the rest of the methods for the simulated data.

The number of changes in state of the indicators was generally similar. The fixed effect variants tended to perform poorly, so using a random effects prior (global adaptation) improved the mixing. The effect of the hierarchical variance is to pull the posteriors for the β_j's towards the right part of the parameter space, so that when $I_j = 0$, β_j is being sampled from close to the correct part of the parameter space. It is interesting that the fixed effect GVS method does not exhibit good mixing.

4.2 Estimation: Simulated Data

The posterior inclusion probabilities of a variable being in a model are plotted against their true values in Fig. 2, and all of the posterior inclusion probabilities are plotted in Fig. 1 of the Supplementary Material. The slope of the fitted line in Fig. 2 indicates how well the method does in distinguishing between important and minor effects, and the position along the x-axis indicates how sensitive the method is to letting smaller effects into the model. The Laplacian models perform poorly, with worse discrimination than the no selection models. The other methods perform similarly to each other, with the exception of the fixed effect GVS with flat priors, which tends to exclude variables, and the Meuwissen & Goddard form of SSVS with informative priors, which behaves inconsistently.

The posterior distribution of the variable with the largest true effect size is shown in Fig. 3. The conditional distributions (i.e. $P(\beta_j \mid I_j = 1, data)$) are similar: the principal difference being the larger uncertainty in the Laplacian estimates. The figures are similar for the second set of priors (see Fig. 2 in Supplementary Material), except that in the case of the random effects SSVS $P(\beta_j \mid I_j = 0, data)$ and $P(\beta_j \mid I_j = 1, data)$ are very similar.

4.3 Estimation: Barley data

The posterior estimates of $P(I_j = 1 \mid data)$ from different methods are shown in Fig. 4. For both set of priors, the No Selection and Laplacian shrinkage methods work badly, showing high marginal posterior occupancy probabilities for all loci. In contrast, the

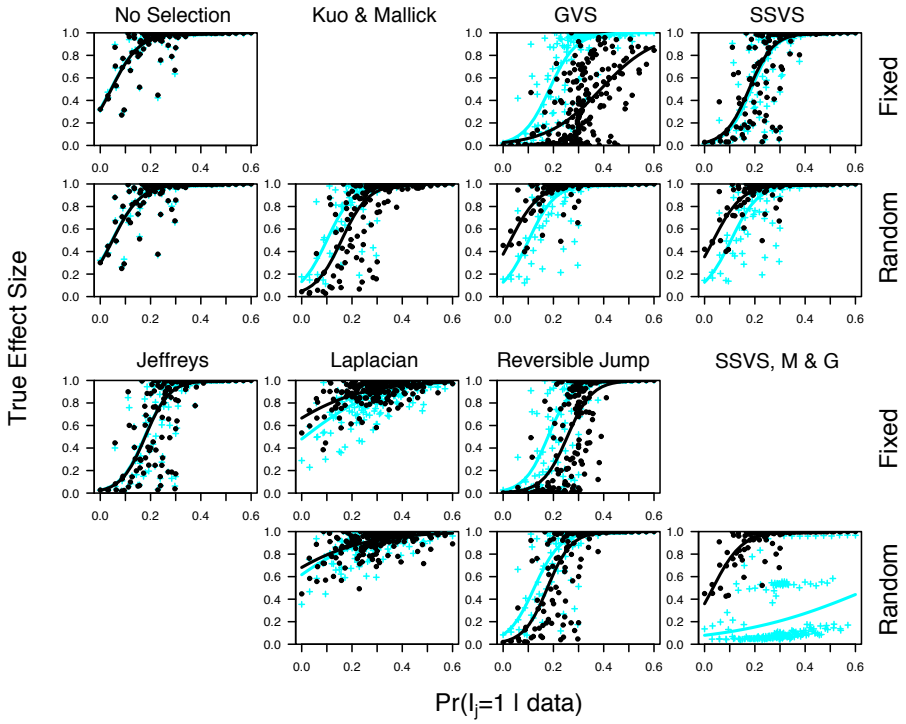

Figure 2: Posterior inclusion probabilities for simulated data plotted against true values of the coefficients. Lines show the fitted curves (quasi-binomial generalized linear model with a logit link). Black and dots: priors set 1, cyan and crosses: priors set 2.

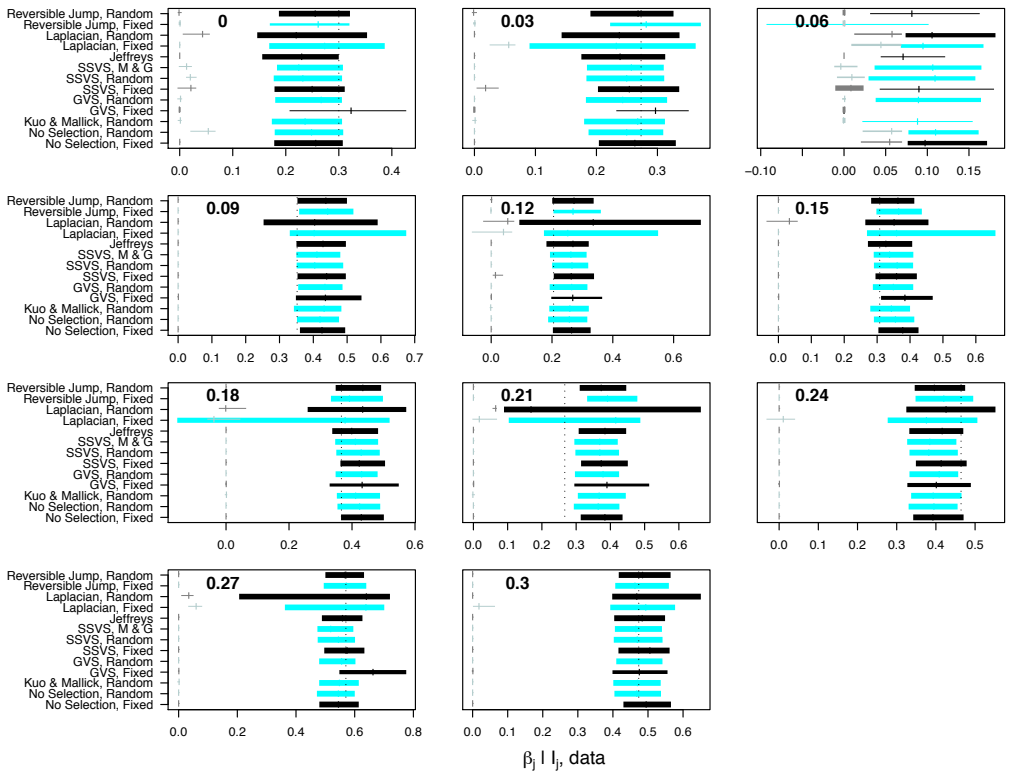

Figure 3: Posterior distributions for regression coefficients, β_j, for the variable with the largest true effect size in the Simulated data, with vague priors. Posterior mode and 70% highest posterior density interval. Black: fixed effect models, cyan: random effect models. Strong colours (i.e. black and red) denote $P(\beta_j \mid I_j = 1, data)$, Lighter colours (i.e. grey and light cyan) denote $P(\beta_j \mid I_j = 0, data)$. The dotted line shows the true effect size of the simulated coefficients. Numbers in plots are b in the simulations of the data: a higher number equates to a larger variation in the true coefficients (see text for details).

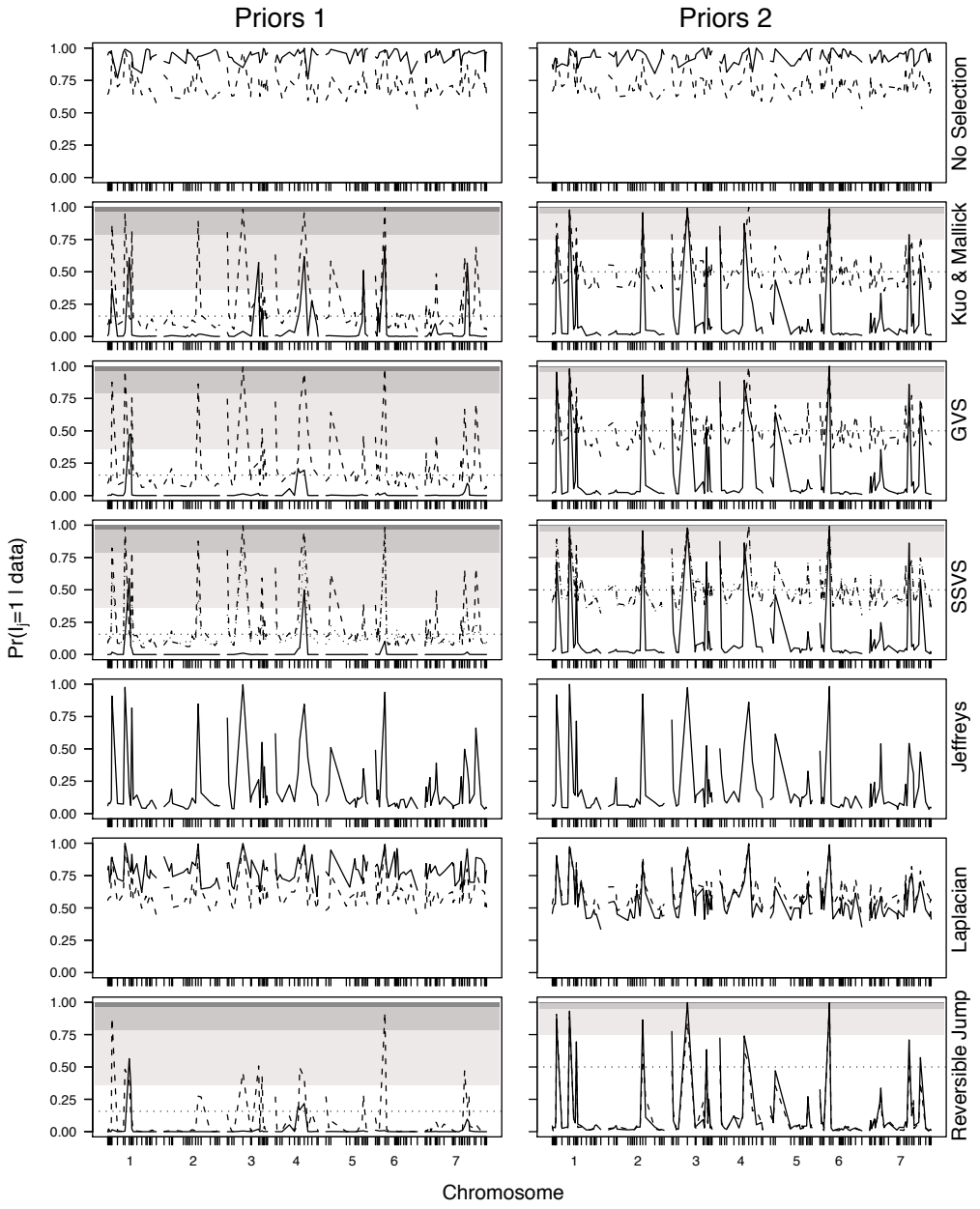

Figure 4: Posterior inclusion probabilities for Barley data. Dotted line: prior probability, light grey region: 3 <Bayes Factor < 20, mid-grey region: 20 <Bayes Factor < 150, dark grey region: 150 <Bayes Factor. Solid line: fixed effect models, dashed line: random effect models.

fixed effect methods and the adaptive shrinkage with Jeffreys' prior tended to give low estimates for $P(I_j = 1 \mid data)$, with only a few variables likely to be in the model, and even these have intermediate probabilities. The random effects methods, wih the exception of the Kuo & Mallick and reversible jump MCMC methods, gave poor separation of the variables, the marginal posterior inclusion probability for variables apparently having no effect was similar to the prior probability: this behavior is clear with the more informative priors too. For the fixed effect variants, the use of informative priors lead to similar posterior distributions, in particular the Kuo & Mallick and reversible jump MCMC methods achieved better identification of variables that should be included in the model.

The posterior distributions of β_j for the 10 loci showing the highest posterior QTL occupancy $P(I_j = 1 \mid data)$ (averaged over the models) are similar (Fig. 5), although the random effects variants shrink the estimates towards zero, as might be expected (these variables are chosen to be extreme, so the common variance shrinks them towards the other variables). The estimates from the analyses with the second set of priors show a similar pattern (results not shown).

4.4 Estimation: Pollution Data

The posterior estimates of $P(I_j = 1 \mid data)$ from the different methods are shown in Fig.6. In general, the fixed effect variants of the methods show a better separation of the variables, more so with the first set of priors. The result are similar to those in the original paper, with Rainfall, January Temperature, Education, percentage non-white and SO2 potential showing large marginal posterior probabilities of variables being in the model: the difference is that July temperature and population density have low posterior probabilities $P(I_j = 1 \mid data)$, whereas they are included in the model after variable selection using Mallow's C_p and ridge regression respectively.

The posterior estimated parameters for Rainfall, January Temperature, percentage non-white and SO2 tend to be fairly similar under both priors (Fig. 3 in Supplementary Material). In the case of the vague priors, the random effect variants of the methods shrink the posterior estimates towards zero.

4.5 Influence of Tuning

Changing the variance of the pseudo-prior in the GVS and the ratio, g, in the M & G model had little effect on marginal posterior probabilities $P(I_j = 1 \mid data)$, but increasing the standard deviation of the spike in the fixed effect SSVS decreased $P(I_j = 1 \mid data)$ values (data not shown). This latter effect is to be expected: if the mass of the likelihood is close to zero, increasing the width of the spike increases the overlap with the likelihood, making it easier for the indicator to flip into state 0.

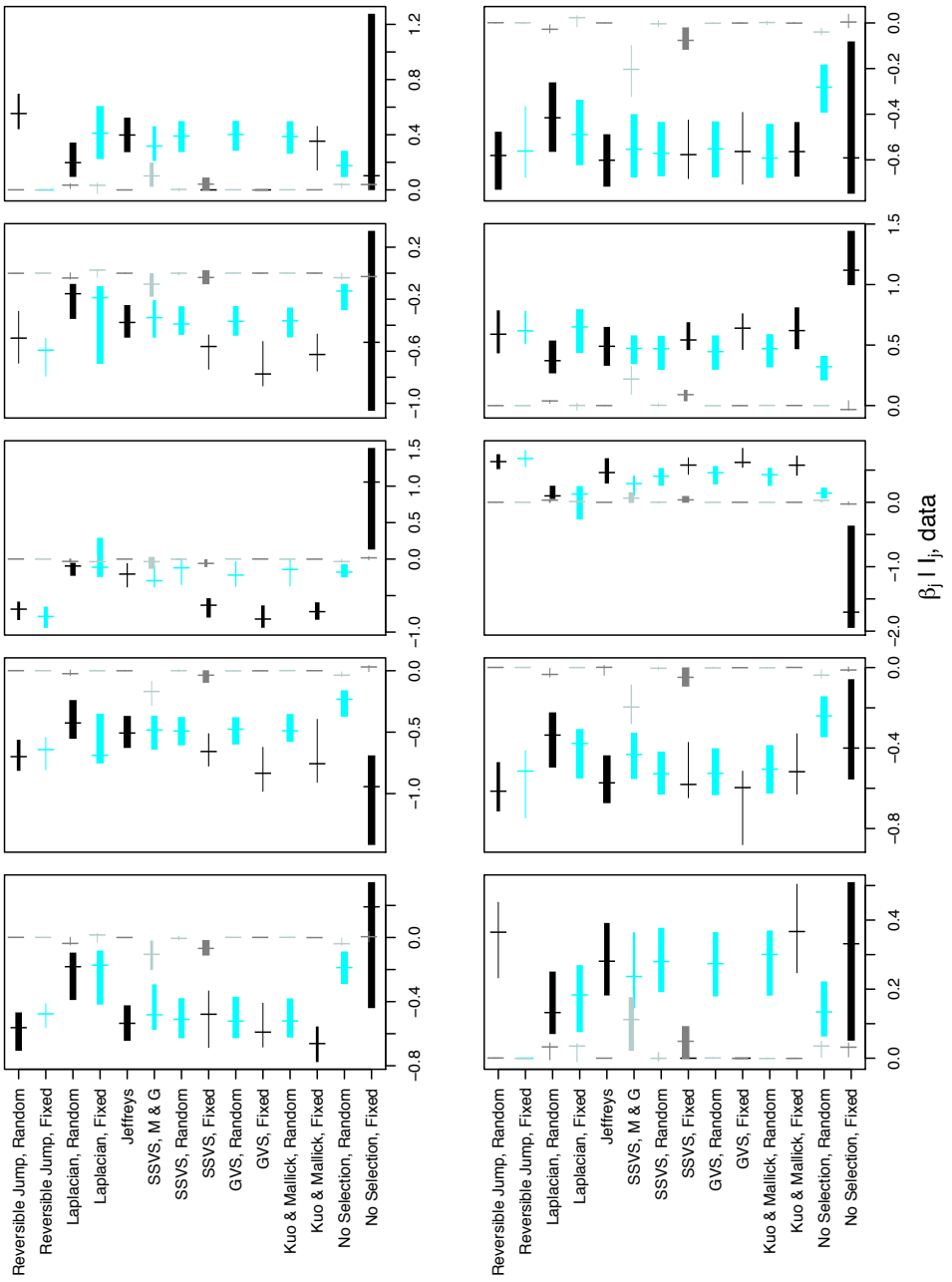

Figure 5: Posterior distributions for regression coefficients, β_j, for the 10 loci (one per panel) showing the highest marginal posterior inclusion probability, $P(I_j = 1 \mid data)$, in the Barley data set, with vague priors. Posterior mode and 70% highest posterior density interval. Black: fixed effect models, cyan: random effect models. Strong colours (i.e. black and cyan) denote $P(\beta_j \mid I_j = 1, data)$, Lighter colours (i.e. grey and light cyan) denote $P(\beta_j \mid I_j = 0, data)$.

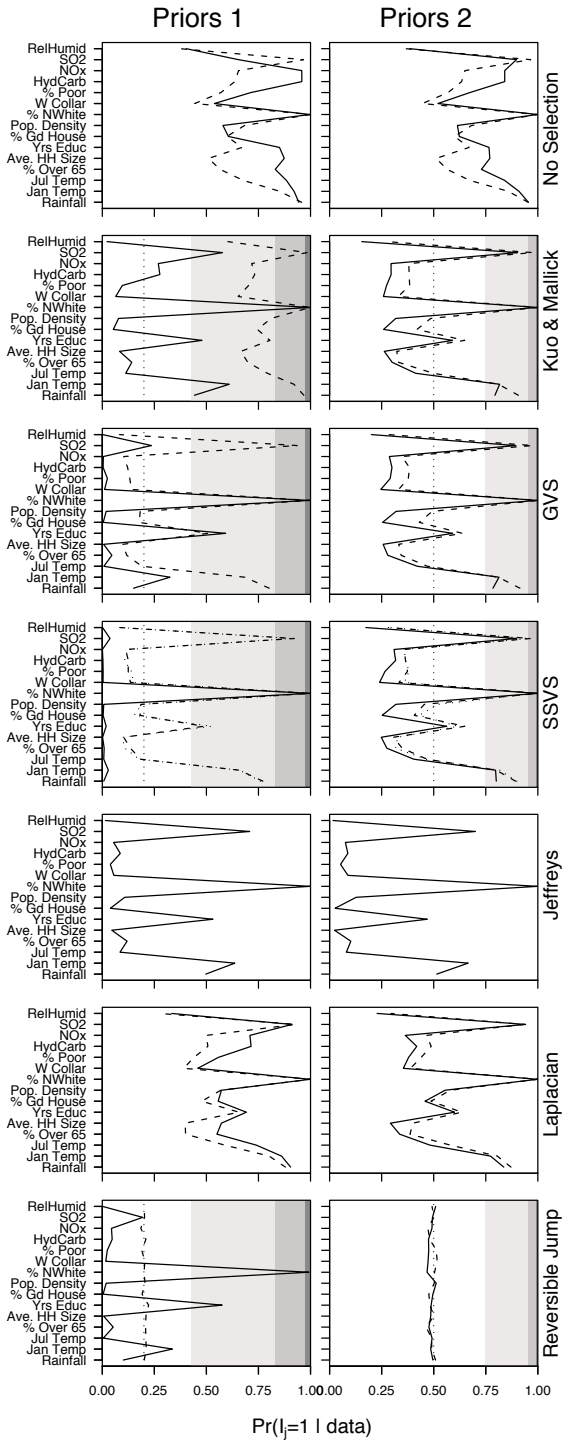

Figure 6: Posterior inclusion probabilities for Pollution data. Dashed line: prior probability, light grey region: 3 <Bayes Factor < 20, mid-grey region: 20 <Bayes Factor < 150, dark grey region: 150 <Bayes Factor. Solid line: fixed effect models, dashed line: random effect models.

Increasing the variance of the GVS pseudo-prior had varying effects: increasing the number of runs for the Pollution data, and the Barley data up to a plateau, had little effect on the Simulations, except for an outlier (Fig. 7, first column). Increasing the width of the spike in the fixed effect SSVS increased the number of runs to a peak, and then decreased it (Fig. 7, second column), with the Pollution data having a second increase. The decrease is related to the change in the posterior probability: as the indicators are more often in state 0, they will flip less often, and hence there will be fewer runs. Because most of the posterior probabilities for the simulated data are close to 1, increasing the width decreases them all towards 0.5, so increases the flipping between states, and hence the number of runs.

Increasing the ratio in the M & G model caused a decline in the number of runs (Fig. 7, third column). If we assume that the width of the "slab" is determined by the data (i.e. it barely changes with g), then the spike much be becoming thinner. This is similar to the fixed effect SSVS: there is less overlap with the likelihood, so it is harder for the indicator to flip between states.

Reducing the prior variance of the regression coefficients in the Laplacian shrinkage method had little effect on the posterior modes (Fig. 8), other than the small prior variances tending to reduce the coefficients towards zero. This effect is clearer when much smaller prior variances are used (data not shown). Unfortunately, it shrinks all of the coefficients, including those that the other methods suggest should be non-zero. Hence, the method gives bad separation of the variables, which is not the desired behaviour.

5 Discussion

The variety of methods available for variable selection is a tribute to the ingenuity of those who have been working on these problems. Each method has its own properties, and it is unlikely that any one will be optimal for all situations. Our conclusions about the methods are summarized in Table 1. Some recommendations can be made on the basis of these conclusions, especially when BUGS is being used. Some caveats are necessary, and these will be explored below.

Firstly, it should be observed that methods based on placing a prior point mass at zero (i.e. the Kuo & Mallick and GVS methods) tended to behave poorly in these tests, being slow and performing no better than the other methods. With non-informative priors, the adaptive shrinkage approach using Jeffreys' prior appeared to work best, being roughly as fast as the other methods, and also mixing well and providing a good separation between variables 'in' and 'out' of the model. If informative priors and a fixed effect model are to be used, then SSVS becomes more attractive: mixing was improved and

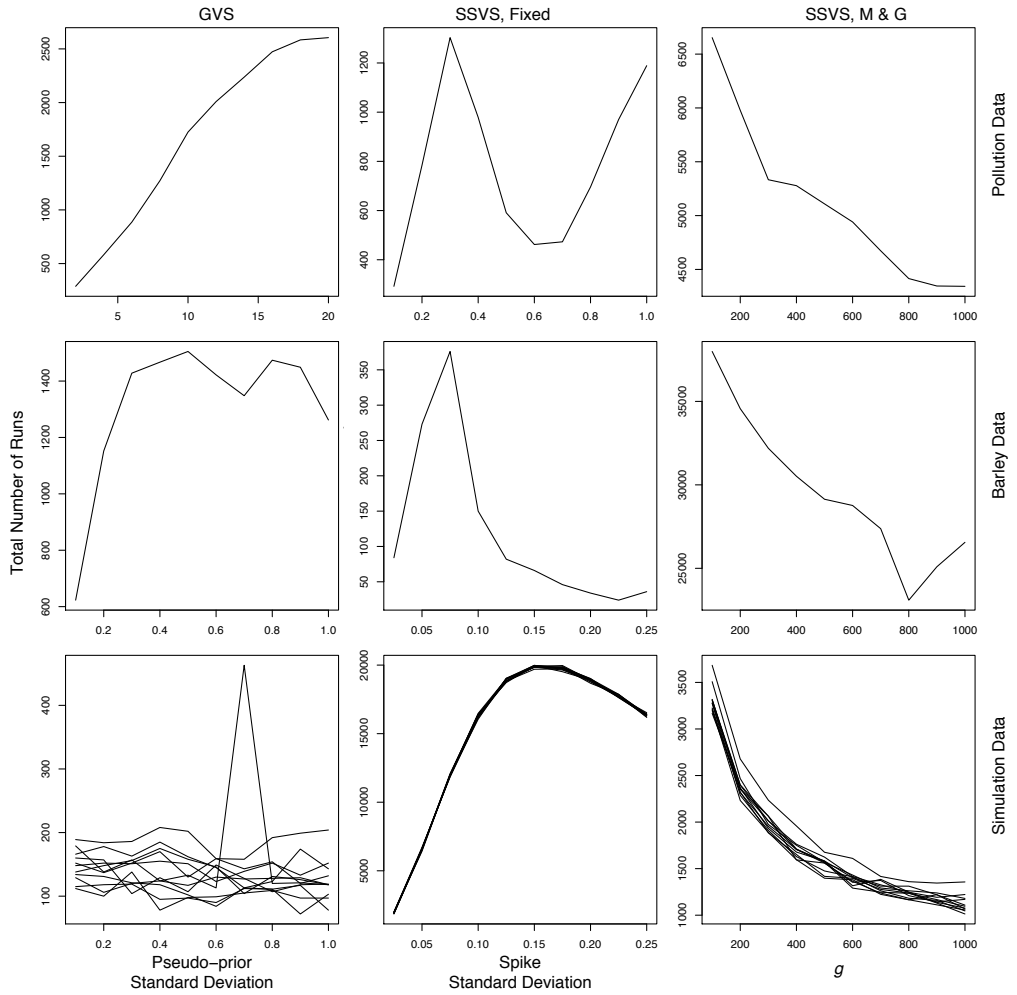

Figure 7: Effect of tuning parameters on the number of runs for three data sets: effect of (1) the width of the pseudo-prior in GVS, (2) the spike width in the fixed effect SSVS, and (3) the ratio of slab to spike variances in the M & G SSVS.

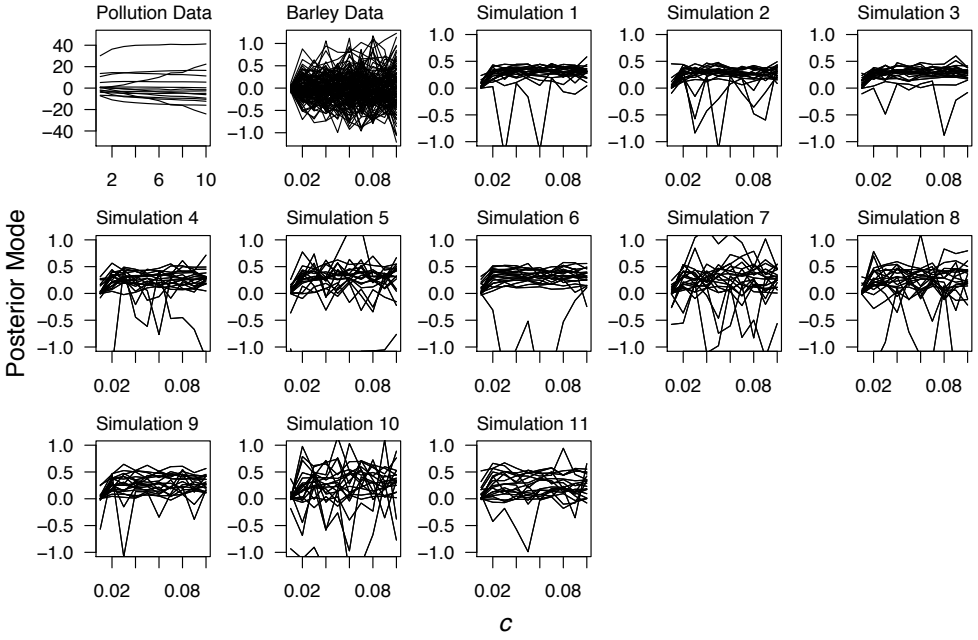

Figure 8: Effect of prior mean (expressed through the "cut-off", c) for Laplacian shrinkage on the posterior modes of β_j.

good separation achieved (as, indeed, it was with the Kuo & Mallick and GVS methods).

The Laplacian method performed poorly, with the results looking similar to the no selection results. This is perhaps not surprising as the prior distributions are similar, without a marked spike. Yi and Xu (2008) investigated the same model, but placed a slightly different prior on μ. Using their prior in our analyses gave similar results to those presented here. The Bayesian LASSO appears not to shrink the variables with small effects, as the "spike" is relatively flat. Yi and Xu (2008) also found this behaviour with a t-distributed prior, and estimating the shape parameter (i.e. the "degrees of freedom"). This suggests that sparsity has to be forced onto a model; the data themselves may not demand it.

For random effect models (utilising global adaptation), all methods give similar results when they work, so the faster speed of SSVS makes it more attractive. The Meuwissen & Goddard implementation can behave poorly (e.g. Fig. 2), and may be sensitive to the prior parameters. One problem in the random effect models is that separation is poorer. This is a result of the posterior variance, τ^2, being pulled down, so that there is less difference between $P(\beta_j \mid I_j = 0, data)$ and $P(\beta_j \mid I_j = 1, data)$ in the region where the MCMC chain is sampling. Hence, if in reality $I_j = 0$, the posterior distribution of I_j is largely determined by the prior, a problem which may be particularly bad when there is little data. One strategy that might help here is to use block-local adaptation: splitting the covariates up into groups which may be expected to react in a similar way, and using a different random effect within each (local) group.

Reversible jump MCMC can run very quickly per iteration but this is offset by the poorer mixing, so that overall it performs similarly to the other methods. Curiously, for the Barley data it estimated a smaller number of variables than the other methods, and seems to perform best (i.e. give a clear separation of $P(I_j = 1 \mid data)$ into values close to 0 or 1 when prior $P(I_j = 1) = 0.5$. Implementing a reversible jump MCMC scheme that analytically integrated out the β_j's would improve the mixing performance of the sampler.

To what extent are the conclusions outlined above generalisable? Firstly, it must be noted that some of the results are specific to BUGS: in particular, the Composite Model Space will be faster than most of the other methods if it is programmed from scratch. Similarly, analytic integration over the β_j's should provide better mixing properties and performance of the methods. Optimisation of code, use of block-updates (e.g. blocking each I_j and β_j together; Meuwissen et al. 2001; Geweke 1996), and local adaptation may also improve performance of the methods. It should also be acknowledged that we considered only three data sets, and that other tests of the methods may lead to different conclusions. This is probably mainly relevant for the comparisons of the mixing performances of the methods, where the data and priors can have a large influence.

The importance of these results to other models also depends on what else are in these other models. The models examined here only had a regression with variable selection, and no other sub-model. For the Barley data there are over 100 variables, so the differences in timings are particularly severe. Hence, in cases where variable selection is carried out for fewer variables, and where the variable selection is part of a larger model, any differences between models may not be of practical significance. This may mean that differences in timings and mixing may be small, so GVS or Kuo & Mallick (for example) will perform just as well as SSVS. One advantage of GVS and Kuo & Mallick is that when $I_i = 0$, $\theta_j = 0$, so no tuning of the width of the spike is required.

The results here suggest that the random effect approach (i.e. global adaptation) shows promise as a method that tunes the model to make variable selection easier, but it perhaps needs further study. As already noted, the posterior probability is approximately equal to the prior probability when there is no evidence that the variable should be included, so Bayes factors should probably be used to present the results. The random effect methods were less effective when there are fewer variables in the model. In this case, the variance of the random effect is being more poorly estimated, and hence the tuning is less accurate, as can be seen in the Pollution example where GVS, SSVS, and reversible jump MCMC gave very high posterior probabilities to all variables, with the distributions of $\beta_j \mid I_j$ being very similar for $I_j = 0$ and $I_j = 1$. It is possible that, in some cases, the slab part of the model will become shrunk to close to zero, in which case the model may become stuck suggesting that many variables are in the model, but all with very small effects. Here we have assumed that p, the prior inclusion probability, was fixed. But another form of global adaptation would be to place a prior on this, and hence estimate a global inclusion probability (i.e. common over all js) (Iswaran and Rao 2005) or analytically integrate it out of the analysis (Smith and Kohn 2002). Our experiments with this idea suggest that it will not perform better than the methods investigated here (results not shown).

Using informative priors and local adaptation will also help the mixing, so it is advisable to consider if they can be elicited. In principle there is considerable flexibility in the priors that could be used, for example Sillanpää and Bhattacharjee (2005, 2006) used the Kuo & Mallick approach, with local adaptation using a Cauchy distribution as a prior for β_j. Because this distribution also has a large amount of probability mass around zero, there was considerable confounding between the parameters (in the manner discussed above). The authors therefore focused on estimating θ_j rather than I_j, and the method is perhaps better seen as a variant of adaptive shrinkage. Even when informative priors cannot be justified from prior knowledge, their use may still be considered as long as the computational advantages sufficiently out-weigh the disadvantages in the bias they induce.

Whilst the use of variable selection can be criticised as being hypothesis testing in a fake beard and glasses, there are still occasions when it can be useful, in particular when the purpose of the analysis is exploratory. The example of QTL analysis (i.e. the Barley

data) is typical here, where there is often little *a priori* reason to expect any particular marker to behave differently. It is then perhaps not surprising that much of the recent development has been in this area, although the wider use of these methods will depend on either software being written specifically for an application, or the porting of these methods into general purpose software such as OpenBUGS. It is pleasing, then, that the simpler methods do seem to work well, suggesting that even if they are not optimal, they are still a useful part of the Bayesian's armoury.

References

Ball, R. D. (2001). "Bayesian methods for quantitative trait loci based on model selection: approximate analysis using Bayesian information criteria." *Genetics*, 159: 1351–1364.

Broman, K. W. and Speed, T. P. (2002). "A model selection approach for identification of quantitative trait loci in experimental crosses." *J. Roy. Stat. Soc. B.*, 64: 641–656.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). "Multivariate Bayesian variable selection and prediction." *J. Roy. Stat. Soc. B.*, 60: 627–641.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, U.S.A.: Springer, 2nd edition.

Carlin, B. R. and Chib, S. (1995). "Bayesian model choice via Markov Chain Monte Carlo methods." *J. Roy. Stat. Soc. B.*, 57: 473–484.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000). "Bayesian variable selection using the Gibbs sampling." In Dey, D. K., Ghosh, S. K., and Mallick, B. K. (eds.), *Generalized linear models: a Bayesian perspective*, 273–286. Marcel Dekker, Inc.,New York.

Dellaportas, P., J.J.Forster, and Ntzoufras, I. (1997). "On Bayesian model and variable selection using MCMC. Technical report." Technical report, Department of Statistics, Athens University of Economics and Business, Athens Greece.

Figueiredo, M. A. T. (2003). "Adaptive sparseness for supervised learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25: 1150–1159.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1: 515–534.

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *J. Am. Stat. Assoc.*, 85: 398–409.

— (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 7: 339–373.

Geweke, J. (1996). "Variable selection and model comparison in regression." In Bernando, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 5*, 609–620. Oxford University Press, Oxford, UK.

Geyer, C. J. (1992). "Practical Markov chain Monte Carlo." *Statist. Sci.*, 7: 473–483.

Godsill, S. J. (2001). "On the relationship between MCMC model uncertainty methods." *J. Comput. Graph. Stat.*, 10: 230–248.

Göring, H. H. H., Terwilliger, J. D., and Blangero, J. (2001). "Large upward bias in estimation of locus-specific effects from genomewide scans." *Am. J. Hum. Genet.*, 69: 1357–1369.

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

Hopert, J. P. and Casella, G. (1996). "The effect of improper priors on Gibbs sampling in hierarchical mixed models." *J. Am. Stat. Assoc.*, 91: 1461–1473.

Hoti, F. and Sillanpää, M. J. (2006). "Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits." *Heredity*, 97: 4–18.

Iswaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *The Annals of Statistics*, 33: 730–773.

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *J. Am. Stat. Assoc.*, 90: 773–795.

Kilpikari, R. and Sillanpää, M. J. (2003). "Bayesian analysis of multilocus association in quantitative and qualitative traits." *Genet. Epidemiol.*, 25: 122–135.

Knapp, S., Bridges, W., and Birkes, D. (1990). "Mapping quantitative trait loci using molecular marker linkage maps." *Theor. Appl. Genet.*, 79: 583–592.

Kuo, L. and Mallick, B. (1998). "Variable selection for regression models." *Sankhya Ser. B*, 60: 65–81.

Lande, R. and Thompson, R. (1990). "Efficiency of marker-assisted selection in the improvement of quantitative traits." *Genetics*, 124: 743–756.

Lunn, D. J., Whittaker, J. C., and Best, N. (2006). "A Bayesian toolkit for genetic association studies." *Genet. Epidemiol.*, 30: 231–247.

McDonald, G. C. and Schwing, R. C. (1973). "Instabilities of regression estimates relating air pollution to mortality." *Technometrics*, 15: 463–482.

Meuwissen, T. H. E. and Goddard, M. E. (2004). "Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data." *Genet. Sel. Evol.*, 36: 261–279.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). "Prediction of total genetic value using genome-wide dense marker map." *Genetics*, 157: 1819–1829.

Miller, A. (2002). *Subset Selection in Regression*. Boca Raton, Florida, U.S.A.: Chapman & Hall/CRC.

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *J. Am. Stat. Assoc.*, 103: 681–686.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2008). *coda: Output analysis and diagnostics for MCMC*. R package version 0.13-3.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2nd edition.

Sen, S. and Churchill, G. A. (2001). "A statistical framework for quantitative trait mapping." *Genetics*, 159: 371–387.

Sillanpää, M., Gasbarra, D., and Arjas, E. (2004). "Comment on "On the Metropolis-Hastings acceptance probability to add or drop a quantitative trait locus in Markov Chain Monte Carlo-based Bayesian analyses"." *Genetics*, 167: 1037.

Sillanpää, M. J. and Arjas, E. (1998). "Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data." *Genetics*, 148: 1373–1388.

Sillanpää, M. J. and Bhattacharjee, M. (2005). "Bayesian association-based fine mapping in small chromosomal segments." *Genetics*, 169: 427–439.

— (2006). "Association mapping of complex trait loci with context-dependent effects and unknown context-variable." *Genetics*, 174: 1597–1611.

Sillanpää, M. J. and Corander, J. (2002). "Model choice in gene mapping: what and why." *Trends Genet.*, 18: 301–307.

Smith, M. and Kohn, R. (2002). "Parsimonious covariance matrix estimation for longitudinal data." *J. Am. Stat. Assoc.*, 97: 1141–1153.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and der Linde, A. V. (2002). "Bayesian measures of model complexity and fit (with discussion)." *J. R. Stat. Soc. B.*, 64: 583–616.

Sturtz, S., Ligges, U., and Gelman, A. (2005). "R2WinBUGS: A Package for Running WinBUGS from R." *Journal of Statistical Software*, 12: 1–16.

ter Braak, C. J. F., Boer, M. P., and Bink, M. C. A. M. (2005). "Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome." *Genetics*, 170: 1435–1438.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). "Making BUGS Open." *R News*, 9: 12–17.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *J. R. Stat. Soc. B.*, 58: 267–288.

Tinker, N. A., Mather, D. E., Rossnagel, B. G., Kasha, K. J., and Keinhofs, A. (1996). "Regions of the genome that affect agronomic performance in two-row barley." *Crop. Sci.*, 36: 1053–1062.

Tipping, M. E. (2004). "Bayesian inference: an introduction to principles and practice in machine learning." In Bousquet, O., von Luxburg, U., and Rätsch, G. (eds.), *Advanced Lectures on Machine Learning*, 41–46. Springer.

Uimari, P. and Hoeschele, I. (1997). "Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algoritm." *Genetics*, 146: 735–743.

Xu, S. (2003). "Estimating polygenic effects using markers of the entire genome." *Genetics*, 163: 789–801.

Yi, N. (2004). "A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci." *Genetics*, 167: 967–975.

Yi, N., George, V., and Allison, D. B. (2003). "Stochastic search variable selection for identifying multiple quantitative trait loci." *Genetics*, 164: 1129–1138.

Yi, N. and Xu, S. (2008). "Bayesian LASSO for quantitative trait loci mapping." *Genetics*, 179: 1045–1055.

Zhang, Y.-M. and Xu, S. (2005). "A penalized maximum likelihood method for estimating epistatic effects of QTL." *Heredity*, 95: 96–104.

Acknowledgments

We would like to thank Nengjun Yi and Roderick D. Ball for their comments on the manuscript. This work was supported by research grants (202324 and 205371) from the Academy of Finland.