Bayesian Variable Selection Using the Gibbs Sampler

Petros Dellaportas
Jonathan J. Forster
Ioannis Ntzoufras

ABSTRACT Specification of the linear predictor for a generalised linear model requires determining which variables to include. We consider Bayesian strategies for performing this variable selection. In particular we focus on approaches based on the Gibbs sampler. Such approaches may be implemented using the publically available software BUGS. We illustrate the methods using a simple example. BUGS code is provided in an appendix.

1 Introduction

In a Bayesian analysis of a generalised linear model, model uncertainty may be incorporated coherently by specifying prior probabilities for plausible models and calculating posterior probabilities using

$$f(m|\boldsymbol{y}) \quad = \quad \frac{f(m)f(\boldsymbol{y}|m)}{\sum\limits_{m \in \mathcal{M}} f(m)f(\boldsymbol{y}|m)}, \qquad m \in \mathcal{M} \tag{1.1}$$

where m denotes the model, $\mathcal{M}$ is the set of all models under consideration, $f(m)$ is the prior probability of model m. The observed data $\boldsymbol{y}$ contribute to the posterior model probabilities through $f(\boldsymbol{y}|m)$, the marginal likelihood calculated using $f(\boldsymbol{y}|m) = \int f(\boldsymbol{y}|m, \boldsymbol{\beta}_m)f(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m$ where $f(\boldsymbol{\beta}_m|m)$ is the conditional prior distribution of $\boldsymbol{\beta}_m$, the model parameters for model m and $f(\boldsymbol{y}|m, \boldsymbol{\beta}_m)$ is the likelihood of the data $\boldsymbol{y}$ under model m.

In particular, the relative probability of two competing models m_1 and

m_2 reduces to

$$\frac{f(m_1|\boldsymbol{y})}{f(m_2|\boldsymbol{y})} = \frac{f(m_1)}{f(m_2)} \frac{\int f(\boldsymbol{y}|m_1,\boldsymbol{\beta}_{m_1})f(\boldsymbol{\beta}_{m_1}|m_1)\, d\boldsymbol{\beta}_{m_1}}{\int f(\boldsymbol{y}|m_2,\boldsymbol{\beta}_{m_2})f(\boldsymbol{\beta}_{m_2}|m_2)\, d\boldsymbol{\beta}_{m_2}} \qquad (1.2)$$

which is the familiar expression relating the posterior and prior odds of two models in terms of the Bayes factor, the second ratio on the right hand side of (1.2).

The principal attractions of this approach are that (1.1) allows the calculation of posterior probabilities of all competing models, regardless of their relative size or structure, and this model uncertainty can be incorporated into any decisions or predictions required (Draper, 1995, gives examples of this).

Generalised linear models are specified by three components, distribution, link and linear predictor. Model uncertainty may concern any of these, and the approach outlined above is flexible enough to deal with this. In this chapter, we shall restrict attention to variable selection problems, where the models concerned differ only in the form of the linear predictor. Suppose that there are p possible covariates which are candidates for inclusion in the linear predictor. Then each $m \in \mathcal{M}$ can be naturally represented by a p-vector $\boldsymbol{\gamma}$ of binary indicator variables determining whether or not a covariate is included in the model, and $\mathcal{M} \subset \{0,1\}^p$. The linear predictor for the generalised linear model determined by $\boldsymbol{\gamma}$ may be written as

$$\boldsymbol{\eta} = \sum_{i=1}^{p} \gamma_i \boldsymbol{X}_i \boldsymbol{\beta}_i \qquad (1.3)$$

where $\boldsymbol{\beta}$ is the 'full' parameter vector with dimension p, and $\boldsymbol{X}_i$ and $\boldsymbol{\beta}_i$ are the design sub-matrix and parameter vector, corresponding to the ith covariate. This specification allows for covariates of dimension greater than 1, for example terms in factorial models.

There has been a great deal of recent interest in Bayesian approaches for identifying promising sets of predictor variables. See for example Brown *et al.*(1998) and Chipman (1996, 1997), Clyde *et al.*(1996), Clyde and DeSimone-Sasinowska (1997), George *et al.*(1996), George and McCulloch (1993, 1996, 1997), Geweke (1996), Hoeting *et al.*(1996), Kuo and Mallick (1998), Mitchell and Beauchamp (1988), Ntzoufras *et al.*(1997), Smith and Kohn (1996) and Wakefield and Bennet (1996).

Most approaches require some kind of analytic, numerical or Monte Carlo approximation because the integrals involved in (1.2) are only analytically tractable in certain restricted examples. A further problem is that the size of the set of possible models $\mathcal{M}$ may be extremely large, so that calculation or approximation of $f(\boldsymbol{y}|m)$ for all $m \in \mathcal{M}$ is very time consuming. One of the most promising approaches has been Markov chain Monte Carlo (MCMC). MCMC methods enable one, in principle, to obtain observations from the joint posterior distribution of $(m, \boldsymbol{\beta}_m)$ and consequently estimate $f(m|\boldsymbol{y})$ and $f(\boldsymbol{\beta}_m|m, \boldsymbol{y})$.

In this chapter we restrict attention to model determination approaches which can be implemented by using one particular MCMC method, the Gibbs sampler. The Gibbs samper is particularly convenient for Bayesian computation in generalised linear models, due to the fact that posterior distributions are generally log-concave (Dellaportas and Smith, 1992). Furthermore, the Gibbs sampler can be implemented in a straightforward manner using the BUGS software (Spiegelhalter *et al.*, 1996a). To facilitate this, we provide BUGS code for various approaches in Appendix A.

The rest of the chapter is organised as follows. Section 2 describes several variable selection strategies that can be implemented using the Gibbs sampler Section 3 contains an illustrative example analysed using BUGS code. We conclude this chapter with a brief discussion in Section 4.

2 Gibbs Sampler Based Variable Selection Strategies

As we are assuming that model uncertainty is restricted to variable selection, m is determined by $\boldsymbol{\gamma}$. We require a MCMC approach for obtaining observations from the joint posterior distribution of $f(m, \boldsymbol{\beta}_m)$. The Gibbs sampler achieves this by generating successively from univariate conditional distributions, so, in principle, the Gibbs sampler is determined by $f(m, \boldsymbol{\beta}_m)$. However, flexibility in the choice of parameter space, likelihood and prior has led to a number of different Gibbs sampler variable selection approaches being proposed.

The first method we shall discuss is a general Gibbs sampler based model determination strategy. The others have been developed more specifically for variable selection problems.

2.1 Carlin and Chib's Method

This method, introduced by Carlin and Chib (1995) is a flexible Gibbs sampling strategy for any situation involving model uncertainty. It proceeds by considering the extended parameter vector $(m, \boldsymbol{\beta}_k; k \in \mathcal{M})$. If a sample can be generated from the joint posterior density for this extended parameter, a sample from the required posterior distribution $f(m, \boldsymbol{\beta}_m)$ can be extracted easily.

A joint prior distribution for m and $(\boldsymbol{\beta}_k; k \in \mathcal{M})$ is required. Here, $(\boldsymbol{\beta}_k; k \in \mathcal{M})$ contains the model parameters for every model in $\mathcal{M}$. Carlin and Chib (1995) specify the joint prior distribution through the marginal prior model probability $f(m)$ and prior density $f(\boldsymbol{\beta}_m|m)$ for each model, as above, together with independent 'pseudoprior' or linking densities $f(\boldsymbol{\beta}_{m'}|m \neq m')$ for each model.

The conditional posterior distributions required for the Gibbs sampler are

$$f(\boldsymbol{\beta}_{m'}|m,\ \{\boldsymbol{\beta}_k : k \in \mathcal{M} \setminus \{m'\}\}, \boldsymbol{y},) \propto \begin{cases} f(\boldsymbol{y}|m, \boldsymbol{\beta}_m) f(\boldsymbol{\beta}_m|m) & m' = m \\ f(\boldsymbol{\beta}_{m'}|m) & m' \neq m \end{cases} \tag{1.4}$$

$$f(m|\{\boldsymbol{\beta}_k : k \in \mathcal{M}\}, \boldsymbol{y}) = \frac{A_m}{\sum\limits_{k \in \mathcal{M}} A_k}. \tag{1.5}$$

where

$$A_m = f(\boldsymbol{y}|m, \boldsymbol{\beta}_m) \prod_{s \in \mathcal{M}} [f(\boldsymbol{\beta}_s|m)] f(m), \quad \forall \ \ m \in \mathcal{M}.$$

Therefore, when $m' = m$, we generate from the usual conditional posterior for model m, and when $m' \neq m$ we generate from the corresponding pseudoprior, $f(\boldsymbol{\beta}_{m'}|m)$. The model indicator m is generated as a discrete random variable using (1.5).

The pseudopriors have no influence on $f(\boldsymbol{\beta}_m|m)$, the marginal posterior distribution of interest. They act as a linking density, and careful choice of pseudoprior is essential, if the Gibbs sampler is to be sufficiently mobile.

Ideally, $f(\boldsymbol{\beta}_{m'}|m \neq m')$ should resemble the marginal posterior distribution $f(\boldsymbol{\beta}_{m'}|m', \boldsymbol{y})$, and Carlin and Chib suggest strategies to achieve this.

The flexibility of this method lies in the facility to specify pseudopriors which help the sampler run efficiently. This may also be perceived as a drawback in problems where there are a large number of models under consideration, such as variable selection involving a moderate number of potential variables. Then, specification of efficient pseudopriors may become too time-consuming. A further drawback of the method is the requirement to generate every $\boldsymbol{\beta}_{m'}$ at each stage of the sampler. (This may be avoided by using a 'Metropolis-Hastings' step to generate m, but is outside the scope of the current chapter; see Dellaportas *et al.*, 1997, for details).

Examples which show how BUGS can be used to perform this method can be found in Spiegelhalter *et al.*(1996b).

2.2 Stochastic Search Variable Selection

Stochastic Search Variable Selection (SSVS) was introduced by George and McCulloch (1993) for linear regression models and has been adapted for more complex models such as pharmacokinetic models (Wakefield and Bennett, 1996), construction of stock portfolios in finance (George and McCulloch, 1996), generalised linear models (George *et al.*, 1996, George and McCulloch, 1997), log-linear models (Ntzoufras *et al.*, 1997) and multivariate regression models (Brown *et al.*, 1998).

The difference between SSVS and other variable selection approaches is that the parameter vector $\boldsymbol{\beta}$ is specified to be of full dimension p under all models, so the linear predictor is

$$\boldsymbol{\eta} = \sum_{i=1}^{p} \boldsymbol{X}_i \boldsymbol{\beta}_i. \tag{1.6}$$

Therefore $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ for all models, where $\boldsymbol{X}$ contains all the potential explanatory variables. The indicator variables γ_i are involved in the modelling process through the prior

$$\boldsymbol{\beta}_i|\gamma_i \sim \gamma_i N(0, c_i^2 \Sigma_i) + (1-\gamma_i) N(0, \Sigma_i) \tag{1.7}$$

for specified c_i and Σ_i. The prior parameters c_i and Σ_i in (1.7) are chosen so that when $\gamma_i = 0$ (covariate is 'absent' from the linear predictor) the prior distribution for $\boldsymbol{\beta}_i$ ensures that $\boldsymbol{\beta}_i$ is constrained to be 'close to $\mathbf{0}$'. When $\gamma_i = 1$ the prior is diffuse, assuming that little prior information is available about $\boldsymbol{\beta}_i$.

The full conditional posterior distributions of $\boldsymbol{\beta}_i$ and γ_i are given by

$$f(\boldsymbol{\beta}_i|\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{\setminus i}) \propto f(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\beta})f(\boldsymbol{\beta}_i|\gamma_i)$$

and

$$\frac{f(\gamma_i=1|\boldsymbol{y},\boldsymbol{\gamma}_{\setminus i},\boldsymbol{\beta})}{f(\gamma_i=0|\boldsymbol{y},\boldsymbol{\gamma}_{\setminus i},\boldsymbol{\beta})} = \frac{f(\boldsymbol{\beta}|\gamma_i=1,\boldsymbol{\gamma}_{\setminus i})}{f(\boldsymbol{\beta}|\gamma_i=0,\boldsymbol{\gamma}_{\setminus i})}\frac{f(\gamma_i=1,\boldsymbol{\gamma}_{\setminus i})}{f(\gamma_i=0,\boldsymbol{\gamma}_{\setminus i})} \tag{1.8}$$

where $\boldsymbol{\gamma}_{\setminus i}$ denotes all terms of $\boldsymbol{\gamma}$ except γ_i.

If we use the prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ defined by (1.7) and assume that $f(\gamma_i=0,\boldsymbol{\gamma}_{\setminus i}) = f(\gamma_i=1,\boldsymbol{\gamma}_{\setminus i})$ for all i, then

$$\frac{f(\gamma_i=1|\boldsymbol{y},\boldsymbol{\gamma}_{\setminus i},\boldsymbol{\beta})}{f(\gamma_i=0|\boldsymbol{y},\boldsymbol{\gamma}_{\setminus i},\boldsymbol{\beta})} = c_i^{-d_i} exp\left(0.5\frac{c_i^2-1}{c_i^2}\boldsymbol{\beta}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\beta}_i\right) \tag{1.9}$$

where d_i is the dimension of $\boldsymbol{\beta}_i$.

The prior for $\boldsymbol{\gamma}$ with each term present or absent independently with probability 1/2 may be considered non-informative in the sense that it gives the same weight to all possible models. George and Foster (1997) argue that this prior can be considered as informative because it puts more weight on models of size close to $p/2$. However, posterior model probabilities are most heavily dependent on the choice of the prior parameters c_i^2 and Σ_i. One way of specifying these is by setting $c_i^2\Sigma_i$ as a diffuse prior (for $\gamma_i = 1$) and then choosing c_i^2 by considering the the value of $|\boldsymbol{\beta}_i|$ at which the densities of the two components of the prior distribution are equal. This can be considered to be the smallest value of $|\boldsymbol{\beta}_i|$ at which the term is considered of practical significance. George and McCulloch (1993) applied this approach. Ntzoufras *et al.*(1997) considered log-linear interaction models where $\boldsymbol{\beta}_i$ terms are multidimensional.

2.3 Unconditional Priors for Variable Selection

Kuo and Mallick (1998) advocated the use of the linear predictor $\boldsymbol{\eta} = \sum_{i=1}^p \gamma_i \boldsymbol{X}_i\boldsymbol{\beta}_i$ introduced in (1.3) for variable selection. They considered a prior distribution $f(\boldsymbol{\beta})$ which is independent of $\boldsymbol{\gamma}$ (and therefore $\mathcal{M}$) so that $f(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{\setminus i},\boldsymbol{\gamma}) = f(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{\setminus i})$

Therefore, the full conditional posterior distributions are given by

$$f(\boldsymbol{\beta}_i|\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{\backslash i}) \propto \begin{cases} f(\boldsymbol{y}|\gamma,\boldsymbol{\beta})f(\boldsymbol{\beta}_i|,\boldsymbol{\beta}_{\backslash i}) & \gamma_i = 1 \\ f(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{\backslash i}) & \gamma_i = 0 \end{cases} \tag{1.10}$$

and

$$\frac{f(\gamma_i=1|\boldsymbol{y},\boldsymbol{\gamma}_{\backslash i},\boldsymbol{\beta})}{f(\gamma_i=0|\boldsymbol{y},\boldsymbol{\gamma}_{\backslash i},\boldsymbol{\beta})} = \frac{f(\boldsymbol{y}|\gamma_i=1,\boldsymbol{\gamma}_{\backslash i},\boldsymbol{\beta})}{f(\boldsymbol{y}|\gamma_i=0,\boldsymbol{\gamma}_{\backslash i},\boldsymbol{\beta})}\frac{f(\gamma_i=1,\boldsymbol{\gamma}_{\backslash i})}{f(\gamma_i=0,\boldsymbol{\gamma}_{\backslash i})}. \tag{1.11}$$

The advantage of the above approach is that it is extremely straightforward. It is only required to specify the usual prior on $\boldsymbol{\beta}$ (for the full model) and the conditional prior distributions $f(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{\backslash i})$ replace the pseudopriors required by Carlin and Chib's method. However, this simplicity may also be a drawback, as there is no flexibility here to alter the method to improve efficiency. In practice, if, for any $\boldsymbol{\beta}_i$, the prior is diffuse compared with the posterior, the method may be inefficient.

2.4 Gibbs Variable Selection

Dellaportas *et al.*(1997) considered a natural hybrid of SSVS and the 'Unconditional Priors' approach of Kuo and Mallick (1998). The linear predictor is assumed to be of the form of (1.3) where , unlike SSVS, variables corresponding to $\gamma_i = 0$ are genuinely excluded from the model. The prior for $(\boldsymbol{\gamma},\boldsymbol{\beta})$ is specified as $f(\boldsymbol{\gamma},\boldsymbol{\beta}) = f(\boldsymbol{\gamma})f(\boldsymbol{\beta}|\boldsymbol{\gamma})$. Consider the partition of $\boldsymbol{\beta}$ into $(\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}})$ corresponding to those components of $\boldsymbol{\beta}$ which are included ($\gamma_i = 1$) or not included ($\gamma_i = 0$) in the model, then the prior $f(\boldsymbol{\beta}|\boldsymbol{\gamma})$ may be partitioned into model prior $f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\gamma)$ and pseudoprior $f(\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\gamma)$.

The full conditional posterior distributions are given by

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}},\gamma,\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta},\gamma)f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\gamma)f(\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\gamma) \tag{1.12}$$

$$f(\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\gamma,\boldsymbol{y}) \propto f(\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\gamma) \tag{1.13}$$

and

$$\frac{f(\gamma_i=1|\boldsymbol{\gamma}_{\backslash i},\boldsymbol{\beta},\boldsymbol{y})}{f(\gamma_i=0|\boldsymbol{\gamma}_{\backslash i},\boldsymbol{\beta},\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_i=1,\boldsymbol{\gamma}_{\backslash i})}{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_i=0,\boldsymbol{\gamma}_{\backslash i})}\,\frac{f(\boldsymbol{\beta}|\gamma_i=1,\boldsymbol{\gamma}_{\backslash i})}{f(\boldsymbol{\beta}|\gamma_i=0,\boldsymbol{\gamma}_{\backslash i})}\,\frac{f(\gamma_i=1,\boldsymbol{\gamma}_{\backslash i})}{f(\gamma_i=0,\boldsymbol{\gamma}_{\backslash i})}. \tag{1.14}$$

This approach is simplified if it is assumed that the prior for $\boldsymbol{\beta}_i$ depends only on γ_i and is given by

$$f(\boldsymbol{\beta}_i|\boldsymbol{\gamma}_i) = \gamma_i N(0, \Sigma_i) + (1-\gamma_i) N(\tilde{\mu}_i, S_i). \tag{1.15}$$

This prior, where $f(\boldsymbol{\beta}_i|\boldsymbol{\gamma}) = f(\boldsymbol{\beta}_i|\gamma_i)$ potentially makes the method less efficient and is most appropriate in examples where $\boldsymbol{X}$ is orthogonal. In prediction, rather than inference about the variables themselves is of primary interest, then $\boldsymbol{X}$ may always be chosen to be orthogonal (see Clyde *et al.*, 1996).

There is a similarity between this prior and the prior used in SSVS. However, here the full conditional posterior distribution is given by

$$f(\boldsymbol{\beta}_i|\boldsymbol{\gamma}, \boldsymbol{\beta}_{\setminus i}, \boldsymbol{y}) \propto \begin{cases} f(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\beta}) N(0, \Sigma_i) & \gamma_i = 1 \\ N(\tilde{\mu}_i, S_i) & \gamma_i = 0 \end{cases}$$

and a clear difference between this and SSVS is that the pseudoprior $f(\boldsymbol{\beta}_i|\gamma_i = 0)$ does not affect the posterior distribution and may be chosen as a 'linking density' to increase the efficiency of the sampler, in the same way as the pseudopriors of Carlin and Chib's method. Possible choices of $\tilde{\mu}_i$ and S_i may be obtained from a pilot run of the full model; see, for example Dellaportas and Forster (1999).

2.5 Summary of Variable Selection Strategies

The similarities and differences between the three Gibbs sampling variable selection methods presented in sections 2.2, 2.3 and 2.4 may easily be summarised by inspecting the conditional probabilities (1.8), (1.11) and, in particular, (1.14).

In SSVS, $f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$ is independent of $\boldsymbol{\gamma}$ and so the first ratio on the right hand side of (1.14) is absent in (1.8). For the 'Unconditional Priors' approach of Kuo and Mallick (1998), the second term on the right hand side of (1.14) is absent in (1.11) as $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are *a priori* independent. For Gibbs Variable Selection, both likelihood and prior appear in the variable selection step. These differences are also evident by looking at the graphical representations of the three methods in Figure 1.

The key differences between the methods (including Carlin and Chib's method) are in their requirements in terms of prior and/or linking densities.

Carlin and Chib's method and GVS both require linking densities whose sole function is to aid the efficiency of the sampler. GVS is less expensive in requirement of pseudopriors, but correspondingly less flexible. The prior parameters in SSVS all have an impact on the posterior, and therefore the densities cannot really be thought of linking densities. The simplest method that described by Kuo and Mallick (1988) does not require one to specify anything other than the usual priors for the model parameters.

3 Illustrative Example: $2 \times 2 \times 2$ Contingency Table

We present an analysis of the data in table 1.1, taken from Healy (1988). This is a three-way table with factors A,B and C. Factor A denotes the condition of the patient (more or less severe), factor B denotes if the patient was accepting antitoxin medication and the (response) factor C denotes whether the patient survived or not.

		Survival(C)	
Condition (A)	Antitoxin (B)	No	Yes
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

TABLE 1.1. Example Dataset.

Purely for illustration purposes, and to present the BUGS code in Appendix A, we model the above data using both log-linear and logistic regression models.

3.1 Log-linear models

We focus attention on hierarchical models including the main effects focussing our interest on associations between model factors and the corresponding interaction terms in the models. Here, $i \in \{1, A, B, C, AB, AC,$

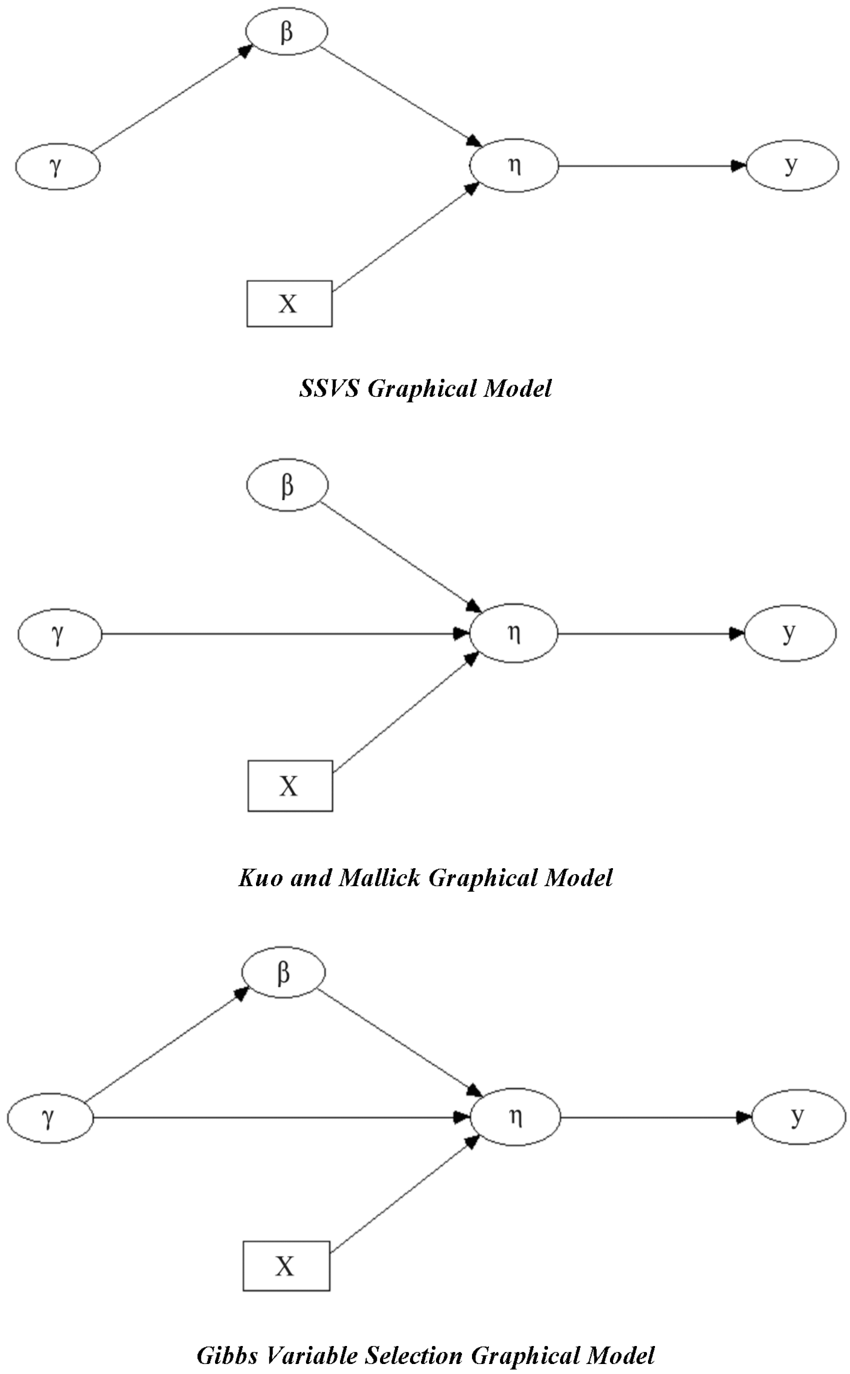

FIGURE 1. Graphical Model Representation for Stochastic Search Variable Selection, Kuo and Mallick Sampler and Gibbs Variable Selection [Squares denote Constants; Circles denote Stochastic Nodes].

$BC, ABC\}$ so $p = 8$. The prior specification for model vector $\boldsymbol{\gamma}$ is $\gamma_i \sim Bernoulli(\pi)$ with $\pi = 1/9$ if $i = ABC$, $\pi = 1$ if $i \in \{1, A, B, C\}$ and $\gamma_i|\gamma_{ABC} \sim Bernoulli(\pi)$ with $\pi = 0.5(1-\gamma_{ABC})+\gamma_{ABC}$ for the two factor interactions ($i \in \{AB, AC, BC\}$). This specification implies that the prior probability of including a two factor interaction in the model is 0.5 if the three factor interaction is excluded from the model and 1 if it is included in the model. Hence the prior probabilities for all 9 possible hierarchical models are 1/9 and and non-hierarchical models are not considered.

For the model coefficients we used the prior specification suggested by Dellaportas and Forster (1999) for log linear models which results in $\Sigma_i = 2$ in (1.15) when the β_i are considered to be the usual 'sum-to-zero' constrained model parameters For SSVS we used $c_i^2\Sigma_i = 2$ and $c_i = 10^3$ in (1.7), as suggested by Ntzoufras *et al.*(1997).

		SSVS	KM	GVS
Models	$A+B+C$	0.1	0.2	0.2
	$AB+C$	0.0	0.1	0.1
	$AC+B$	25.1	25.7	25.6
	$BC+A$	0.3	0.6	0.6
	$AB+AC$	7.9	7.5	7.3
	$AB+BC$	0.1	0.2	0.2
	$AC+BC$	58.9	58.4	58.9
	$AB+BC+CA$	6.4	6.6	6.4
	ABC	1.0	0.8	0.6

TABLE 1.2. Posterior model probabilities (%) for log-linear models. SSVS: Stochastic Search Variable Selection; KM: Kuo and Mallick's Unconditional Priors approach; GVS: Gibbs Variable Selection.

The results are based on 100,000 iterations for Gibbs variable selection and Kuo and Mallick's method, and 400,000 iterations for SSVS which seemed to be less efficient. For all methods we discarded 10,000 iterations as a burn-in period. The pseudoprior densities for Gibbs variable selection were constructed from the sample moments of a pilot run of the full model of size 1,000 iterations. All three methods give similar results supporting the same models with very similar posterior probabilities.

3.2 Logistic regression models

When we consider binomial logistic regression models for response variable C and explanatory factors A and B, there are 5 possible nested models, 1, A, B, $A+B$ and AB. Priors are specified by setting $c_i^2\Sigma_i = 4 \times 2$ in (1.7) and $\Sigma_i = 4 \times 2$ in (1.15) which is equivalent to the prior used above for log-linear model selection. The pseudoprior parameters were specified as before, through a pilot chain, and finally we set $\gamma_{ABC} \sim Bernoulli(1/5)$ and $\gamma_i|\gamma_{AB} \sim Bernoulli(\pi)$, with $\pi = 0.5(1-\gamma_{AB}) + \gamma_{AB}$ for $i \in \{A, B\}$. The resulting prior probabilities for all models are 1/5. The results in table (1.3) are based on 500,000 iterations for SSVS and Kuo and Mallick's method and 100,000 iterations for Gibbs variable selection, with burn-in period of 10,000 iterations. Again, the results are very similar, although Gibbs variable selection seemed to be most efficient.

The equivalent log-linear models in Table 1.2 are those which include the AB term, so the results can be seen to be in good agreement.

		SSVS	KM	GVS
Models	1	0.2	0.5	0.5
	A	48.0	49.2	49.3
	B	1.0	1.2	1.2
	$A+B$	45.3	44.0	43.9
	AB	5.5	5.2	5.1

TABLE 1.3. Posterior model probabilities (%) for logistic regression. SSVS: Stochastic Search Variable Selection; KM: Kuo and Mallick's Unconditional Priors approach; GVS: Gibbs Variable Selection.

4 Discussion

We have reviewed a number of Bayesian variable selection strategies based on the Gibbs sampler. Their major practical advantage is that they can be easily applied with a Gibbs sampling software such as BUGS.

It is impossible to provide a general recommendation for a method of computation for a class of problems as large as variable selection in generalised linear models. The methods we have discussed range from the 'Unconditional Priors approach' which is extremely easy to implement, but may be insufficiently flexible for many practical problems, to the approach of Carlin and Chib, which is very flexible, but requires a lot of careful specification.

We have only discussed methods based on the Gibbs sampler. Of course other extremely flexible MCMC methods exist, such the reversible jump approach introduced by Green (1996). All MCMC methods require careful implementation and monitoring, and other approaches should also be considered. For many model selection problems involving generalised linear models, an alternative approach is through asymptotic approximation. Raftery (1996) has provided a series of Splus routines for this kind of calculation. Such methods can be used in conjunction with the Gibbs sampler approaches discussed here.

Any Bayesian model selection requires careful attention to prior specification. For discussion of elicitation of prior distributions for variable selection, see Garthwaite and Dickey (1992) and Ibrahim and Chen (1998).

5 References

[1] Agresti, A. (1990), *Categorical Data Analysis*, John Wiley and Sons, USA.

[2] Brown, P.J., Vannucci, M. and Fearn, T. (1998), "Multivariate Bayesian Variable Selection and Prediction", *Journal of Royal Statistical Society, B, 60*, 627–641.

[3] Carlin, B.P. and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods", *Journal of Royal Statistical Society, B, 57*, 473–484.

[4] Chipman, H. (1996), "Bayesian Variable Selection with Related Predictors", *Canadian Journal of Statistics, 24*, 17–36.

[5] Chipman, H., Hamada, M., Wu, C.F.J. (1997), "A Bayesian Variable-Selection Approach for Analysing Designed Experiments with Complex Aliasing", *Technometrics, 39*, 372–381.

[6] Clyde, M., DeSimone, H. and Parmigiani, G. (1996), “Prediction via Orthognalized Model Mixing”, Journal of the American Statistical Association, 91, 1197–1208.

[7] Clyde, M. and DeSimone-Sasinowska, H. (1997), “Accounting for Model Uncertainty in Poisson Regression Models: Does Particulate Matter?”, *Technical Report*, Institute of Statistics and Desicion Sciences, Duke University, USA.

[8] Dellaportas, P. and Forster, J.J. (1999), “Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models”, *Biometrika*, to appear.

[9] Dellaportas, P., Forster, J.J. and Ntzoufras, I.(1997), “On Bayesian Model and Variable Selection Using MCMC”, *Technical Report*, Department of Statistics, Athens University of Economics and Business, Greece.

[10] Draper, D. (1995), “Assesment and Propogation of Model Uncertainty” (with discussion), *Journal of the Royal Statistical Society, B, 57*, 45–97.

[11] Garthwaite, P.H. and Dickey, J.M. (1992), “Elicitation of Prior Distributions for Variable-Selection Problems in Regression”, *The Annals of Statistics, 20*, 1697–1719.

[12] George, E.I. and Foster, D.P. (1997), “Calibration and Empirical Bayes Variable Selection”, *Technical Report*, University of Texas at Austin and University of Pennsylvania, USA.

[13] George, E.I. and McCulloch, R.E. (1993), “Variable Selection via Gibbs Sampling”, *Journal of the American Statistical Association, 88*, 881–889.

[14] George, E.I. and McCulloch, R.E. (1996), “Stochastic Search Variable Selection”, *Markov Chain Monte Carlo in Practice*, eds. W.R.Gilks , S.Richardson and D.J.Spiegelhalter, Chapman and Hall, London, UK, 203–214.

[15] George, E.I., McCulloch, R.E. and Tsay R.S. (1996), "Two Approaches for Bayesian Model Selection with Applications", *Bayesian Analysis in Statistics and Econometrics*, eds. D.A.Berry, M.Chaloner and J.K.Geweke, John Wiley and Sons, New York, USA, 339–348.

[16] George, E.I. and McCulloch, R.E. (1997), "Approaches for Bayesian Variable Selection", *Statistica Sinica, 7*, 339–373.

[17] Geweke, J. (1996), "Variable Selection and Model Comparison in Regression" *Bayesian Statistics 5*, eds. J.M.Bernardo, J.O.Berger, A.P.Dawid and A.F.M.Smith, Claredon Press, Oxford, UK, 609–620.

[18] Green, P.J. (1996), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination", *Biometrika, 82*, 711–732.

[19] Healy, M.J.R. (1988), *Glim: An Introduction*, Claredon Press, Oxford, UK.

[20] Hoeting, J.A., Madigan, D., and Raftery, A.E. (1996), "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression", *Journal of Computational Statistics and Data Analysis, 22*, 251-270.

[21] Ibrahim, J.G. and Chen, M.H. (1998), "Prior Elicitation and Variable Selection for Generalised Mixed Models" *Generalized Linear Models: A Bayesian Perspective*, eds. D.K. Dey, S. Ghosh and B. Mallick, Marcel Dekker Publications.

[22] Kuo, L. and Mallick, B. (1998), "Variable Selection for Regression models", *Sankhya, B, 60*, Part 1, 65–81.

[23] Madigan, D. and Raftery, A.E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window", *Journal of the American Statistical Association, 89*, 1535–1546.

[24] Mitchell, T.J. and Beauchamp, J.J. (1988), "Bayesian Variable Selection in Linear Regression", *Journal of the American Statistical Association, 83*, 1023–1036.

[25] Ntzoufras, I., Forster, J.J. and Dellaportas, P. (1997), "Stochastic Search Variable Selection for Log-linear Models", *Technical Report*, Faculty of Mathematics, Southampton University, UK.

[26] Raftery, A.E. (1996), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models", *Biometrika, 83*, 251–266.

[27] Smith M. and Kohn R. (1996), "Nonparametric Regression Using Bayesian Variable Selection", *Journal of Econometrics, 75*, 317–343.

[28] Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W.(1996a), *BUGS 0.5: Bayesian Inference Using Gibbs Sampling Manual*, MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

[29] Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W.(1996b), *BUGS 0.5: Examples Volume 2*, MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

[30] Wakefield, J. and Bennett, J. (1996), "The Bayesian modelling of Covariates for Population Pharmacokinetic Models", *Journal of the American Statistical Association, 91*, 917–927.

6 Appendix: BUGS CODES

Code and data files are freely available in the web adress *http://www.stat-athens.aueb.gr/~jbn/* or by electronic mail request.

6.1 Code for Log-linear Models for 2^3 Contingency Table

```
model loglinear;
#
#       2x2x2 LOG-LINEAR VARIABLE SELECTION WITH BUGS
#       (c) OCTOBER 1996 FIRST VERSION
#       (c) OCTOBER 1997 FINAL VERSION
#           WRITTEN BY IOANNIS NTZOUFRAS
#       ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
#
#           SSVS: Stochastic Search Variable Selection
#           KM  : Kuo and Mallick Gibbs sampler
#           GVS : Gibbs Variable Selection
#
const
   N = 8;  # number of Poisson cells
```

```
var
        include,    # conditional prior probabability for gi
        pmdl[9],    # model indicator vector
        mdl,        # code of model
        b[N],       # model coefficients
        mean[N],    # mean used in pseudoprior   (GVS only)
        se[N],      # st.dev. used in pseudoprior(GVS only)
        bpriorm[N], # prior mean for b depanding on g
        tau[N],     # model coefficients precision
#       c,          # precision multiplicator    (SSVS only)
        x[N,N],     # design matrix
        z[N,N],     # matrix used in likelhood
        n[N],       # Poisson cells
        lambda[N],  # Poisson mean for each cell
        g[N];       # term indicator vector
data n,x in "ex1log.dat", mean, se in 'prop1ll.dat';
inits in "ex1ll.in";
{
#       c<-1000.0 # SSVS only
#
#       calculation of the z matrix used in likelihood
        for (i in 1:N) { for (j in 1:N) {
               z[i,j]<-x[i,j]*b[j]*g[j]   # For GVS/KM
#               z[i,j]<-x[i,j]*b[j];      # For SSVS
               }}
#
#       model configuration
        for (i in 1:N) {
                log(lambda[i])<-sum(z[i,]);
                n[i]~dpois(lambda[i])    }
#       defining model code
#       0 for [A][B][C], 1 for [AB][C],      2 for [AC][B],
#       3 for [AB][AC],  4 for [BC][A],      5 for [AB][BC],
#       6 for [AC][BC],  7 for [AB][BC][CA],15 for [ABC].
#
        mdl<-g[5]+2*g[6]+4*g[7]+8*g[8];
        for (i in 0:7) { pmdl[i+1]<-equals(mdl,i) }
        pmdl[9]<-equals(mdl,15)
#
#       Prior for b model coefficient
        tau[1]<-0.1;
        bpriorm[1]<-0.0;
        b[1]~dnorm(bpriorm[1],tau[1]);
        for (i in 2:N) {
#
#                    GVS using se,mean from pilot run
#               ------------------------------------------
                tau[i]<-g[i]/2+(1-g[i])/(se[i]*se[i]);
                bpriorm[i]<-mean[i]*(1-g[i]);
#
#               Kuo and Mallick (prior indepedent of g[i])
#               ------------------------------------------
#                tau[i]<-1/2;
#                bpriorm[i]<-0.0;
#
#
#                           SSVS PRIOR SET-UP
#               ------------------------------------------
#                tau[i]<-pow(c,2-2*g[i])/2;
```

```
#                bpriorm[i]<-0.0;
#
                b[i]~dnorm(bpriorm[i],tau[i]);
        }
#
#       defining prior information for gi in such way that
#       allow only hierarhical models with equal probability.
#
        include<-(1-g[8])*0.5+g[8]*1.0;
        g[8]~dbern(0.1111111);
        g[7]~dbern(include);
        g[6]~dbern(include);
        g[5]~dbern(include);
        for (i in 1:4) { g[i]~dbern(1.0)}}
```

6.2 Code for Logistic Models with 2 Binary Explanatory Factors

```
model Binomial;
#
#       LOGISTIC REGRESSION VARIABLE SELECTION WITH BUGS
#       (c) OCTOBER 1996 FIRST VERSION
#       (c) OCTOBER 1997 FINAL VERSION
#           WRITTEN BY IOANNIS NTZOUFRAS
#       ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
#
#           SSVS: Stochastic Search Variable Selection
#           KM  : Kuo and Mallick Gibbs sampler
#           GVS : Gibbs Variable Selection
#
const
   N = 4;  # number of binomial experiments
var
      include,   # conditional prior probabability for gi
      pmdl[5],   # model indicator vector
      mdl,       # code of model
      b[N],      # model coefficients
      mean[N],   # mean used in pseudoprior    (GVS only)
      se[N],     # st.dev, used in pseudoprior (GVS only)
      bpriorm[N],# prior mean for b depanding on g
      tau[N],    # model coefficients precision
#     c,         # precision multiplicator    (SSVS only)
      x[N,N],    # design matrix
      z[N,N],    # matrix used in likelhood
      r[N],      # number of successes in binomial
      n[N],      # total number of observations for binomial
      p[N],      # probability of success for binomial model
      g[N];      # term indicator vector
data r,n,x in "ex1logit.dat", mean, se in 'prop1.dat';
inits in "ex1.in";
{
#        c<-1000 # SSVS only
#
#       calculation of the z matrix used in likelihood
        for (i in 1:N) { for (j in 1:N) {
                z[i,j]<-x[i,j]*b[j]*g[j]  # for GVS
#               z[i,j]<-x[i,j]*b[j];      # for SSVS
                }}
#
#       model configuration
```

```
        for (i in 1:N) {
                r[i]~dbin(p[i],n[i]);
                logit(p[i])<-sum(z[i,]) }
#       defining model code
#       0 constant, 1 for [A], 2 for [B],
#       3 for [A][B], and 6 for [AB]
#
        mdl<-g[2]+2*g[3]+3*g[4];
        pmdl[1]<-equals(mdl,0)
        pmdl[2]<-equals(mdl,1)
        pmdl[3]<-equals(mdl,2)
        pmdl[4]<-equals(mdl,3)
        pmdl[5]<-equals(mdl,6)
#
#       Prior for b model coefficient
        tau[1]<-0.1;
        bpriorm[1]<-0.0;
        b[1]~dnorm(bpriorm[1],tau[1]);
        for (i in 2:N) {
#
#               GVS using se,mean from pilot run
#               --------------------------------
#
                tau[i]<-g[i]/8+(1-g[i])/(se[i]*se[i]);
                bpriorm[i]<-mean[i]*(1-g[i]);
#
#               Kuo and Mallick proposal is indedent of g[i]
#               --------------------------------------------
#
#                tau[i]<-1/8;
#                bpriorm[i]<-0.0;
#
#                          SSVS PRIOR SET-UP
#               ------------------------------------------
#                tau[i]<-pow(c,2-2*g[i])/8;
#                bpriorm[i]<-0.0;
#
                b[i]~dnorm(bpriorm[i],tau[i]);
        }
#
#       defining prior information for gi in such way that
#       allow only hierarhical models with 0.2 probability.
#
        g[4]~dbern(0.2);
        include<-(1-g[4])*0.5+g[4]*1.0
        g[2]~dbern(include);
        g[3]~dbern(include);
        g[1]~dbern(1.0)               }
```