ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS
DEPARTMENT OF STATISTICS

Technical Report 64

1999, May: Athens, Greece

Gibbs Variable Selection using BUGS

Ioannis Ntzoufras

Department of Statistics, Athens University of Economics and Business, 12 Kodringtonos, 10434, Athens, Greece.

Summary

In this paper we discuss and present in detail the implementation of Gibbs variable selection as defined by Dellaportas *et al.* (1998) using BUGS. We illustrate the application of Gibbs variable selection using a variety of problems including linear regression, log-linear and binomial response models.

Keywords: Logistic regression; Linear regression; MCMC; Model selection; Random effects.

1 Introduction

Bayesian model selection involves the calculation of posterior model probabilities $f(m|\boldsymbol{y})$. Intractabilities in the calculation of the integrals involved in the posterior model probabilities have been overcomed with the construction of efficient MCMC algorithms for model and variable selection problems.

A variety of MCMC methods have been proposed for variable selection including George and McCulloch's (1993) 'Stochastic Search Variable Selection' (SSVS), reversible jump Metropolis by Green (1995), Carlin and Chib (1995) method, Kuo and Mallick (1998) Gibbs sampler and Gibbs variable selection by Dellaportas *et al.* (1998).

The primary aim of this paper is to clearly illustrate how we can utilize BUGS (Spiegelhalter *et al.* , 1996a) for the implementation of variable selection methods. We concentrate on Gibbs variable selection introduced by Dellaportas *et al.* (1998) with independent prior distributions but extension to SSVS and Kuo and Mallick samplers is straightforward.

The paper is organised into three sections additional to this introductory one. Section 2 describes the general Gibbs variable selection algorithm as introduced by Dellaportas *et al.* (1998), Section 3 provides detailed guidance for implementation in BUGS and finally Section 4 presents three illustrated examples.

2 Gibbs Variable Selection

Many statistical models may be represented naturally as $(s, \boldsymbol{\gamma}) \in \mathcal{S} \times \{0,1\}^p$, where the indicator vector $\boldsymbol{\gamma}$ identifies which of the p possible sets of covariates are present in the model and s denotes other structural properties of the model. For example, for a generalised linear model, s may describe the distribution, link function and variance function, and the linear predictor may be written as

$$\boldsymbol{\eta} = \sum_{j=1}^{p} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j \qquad (1)$$

where $\boldsymbol{X}_j$ is the design matrix and $\boldsymbol{\beta}_j$ the parameter vector related to the jth term. In the following, we restrict attention to variable selection aspects assuming that s is known, or dealt with in another way and therefore we substitute the model indicator m with $\boldsymbol{\gamma}$.

The model parameter vector $\boldsymbol{\beta}$ is partitioned into two vectors $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ corresponding to parameters of variables included or excluded from the model. Then the full conditional posterior distributions are given by

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}},\boldsymbol{\gamma},\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma})f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})f(\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}) \qquad (2)$$

$$f(\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma},\boldsymbol{y}) \propto f(\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}) \qquad (3)$$

and

$$f(\gamma_j|\boldsymbol{\gamma}_{\backslash j},\boldsymbol{\beta},\boldsymbol{y}) = Bernoulli\left(\frac{O_j}{1+O_j}\right) \qquad (4)$$

with

$$O_j = \frac{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_j=1,\boldsymbol{\gamma}_{\backslash j})}{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_j=0,\boldsymbol{\gamma}_{\backslash j})}\,\frac{f(\boldsymbol{\beta}|\gamma_j=1,\boldsymbol{\gamma}_{\backslash j})}{f(\boldsymbol{\beta}|\gamma_j=0,\boldsymbol{\gamma}_{\backslash j})}\,\frac{f(\gamma_j=1,\boldsymbol{\gamma}_{\backslash j})}{f(\gamma_j=0,\boldsymbol{\gamma}_{\backslash j})}. \qquad (5)$$

It is clear that the components of the vector $\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}$ do not affect the model likelihood and therefore are sampled directly from the distribution $f(\boldsymbol{\beta}_{\backslash\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma})$. Such densities are called pseudopriors since they do not carry any further information for the model and their specification only influences the performance of the MCMC algorithm.

The selection of priors and pseudopriors is a very important aspect in model selection. Here we briefly present the simplest approach of independent prior and pseudoprior densities given by

$$f(\boldsymbol{\beta}_j|\boldsymbol{\gamma}_j) = (1-\gamma_j)f(\boldsymbol{\beta}_j|\gamma_j=0)+\gamma_j f(\boldsymbol{\beta}_j|\gamma_j=1). \qquad (6)$$

Note that the above prior can be used in any model selection problem if we orthogonalize the data matrix and then perform model choice using the new transformed data (see Clyde *et al.* , 1996).

The simplified prior (6) and a model formulation such as (1), result in the following full conditional posterior

$$f(\boldsymbol{\beta}_j|\boldsymbol{\gamma},\boldsymbol{\beta}_{\backslash j},\boldsymbol{y}) \propto \begin{cases} f(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\beta})f(\boldsymbol{\beta}_j|\gamma_j=1) & \gamma_j=1 \\ f(\boldsymbol{\beta}_j|\gamma_j=0) & \gamma_j=0 \end{cases} \qquad (7)$$

indicating that the pseudoprior $f(\boldsymbol{\beta}_j|\gamma_j=0)$ does not affect the posterior distribution of each model coefficient.

Similarly to George and McCulloch (1993), we may use a mixture of Normal distribution for model parameters, that is

$$f(\boldsymbol{\beta_j}|\gamma_j=1) \equiv N(0,\Sigma_j) \;\text{ and }\; f(\boldsymbol{\beta_j}|\gamma_j=0) \equiv N(\bar{\mu}_j,S_j). \qquad (8)$$

The hyperparameters $\bar{\mu}_j$ and S_j are parameters of the pseudoprior distribution; therefore their choice is only relevant to the behaviour of the MCMC chain and do not affect the posterior distribution. Ideal choices of these parameters are the maximum likelihood or pilot run estimators of the full model (as, for example, in Dellaportas and Forster, 1999). However, in the experimental process, we noted that an automatic selection of $\bar{\mu}_j = 0$ and

$S_j = \Sigma_j / k^2$ with $k = 10$ has also been proven an adequate choice. This 'automatic selection' uses the properties of the prior distributions with 'large' and 'small' variance introduced in SSVS by George and McCulloch (1993). The parameter k is now only a pseudoprior parameter and therefore it does not affect the posterior distribution. Suitable calibration of this parameter assists the chain to move better (or worse) between different models. In certain cases, defining the pseudoprior parameters from a pilot run of the full model may result in 'badly' proposed values for most models and considerably slow down the Gibbs sampler. In these situations the automatic selection proposed above may work better than the pilot run estimates.

For contingency tables, a prior is proposed by Dellaportas and Forster (1999). Additionally, equivalent priors may be adopted in logistic regression models with categorical explanatory variables (see Dellaportas *et al.* 1999). For generalized linear models Raftery (1996) has proposed to select the prior covariance matrix using elements from the data matrix multiplied by a hyperparameter. The latter was selected in such way that the effect of the prior distribution on the posterior odds becomes minimal. George and Foster (1997) use empirical Bayes criteria to define both priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. O'Hagan (1995) tried to overcome the prior selection problem in model choice using fractional Bayes factor. On the other hand, Berger and Pericchi (1996) introduced intrinsic priors and intrinsic Bayes factor using a minimal sample for prior specification and the rest dataset for model choice.

When no restrictions on the model space are imposed then a common prior for the term indicators γ_j is $f(\gamma_j) = Bernoulli(1/2)$, whereas in other cases (e.g. hierarchical or graphical log-linear models) it is required that $f(\gamma_j | \boldsymbol{\gamma}_{\setminus j})$ depends on $\gamma_{\setminus j}$; for more details see Chipman (1996) and Section 3.3.

Other Gibbs samplers for model selection have also been proposed by George and McCulloch (1993), Carlin and Chib (1995) and Kuo and Mallick (1998). Detailed comparison and discussion of the above methods is given by Dellaportas *et al.* (1998, 1999). Implementation of Carlin and Chib methodology in BUGS is illustrated by Spiegelhater *et al.* (1996c, page 47) while a simple example of the other Gibbs methods is provided by Dellaportas *et al.* (1999).

3 Applying Gibbs Variable Selection Using BUGS

In this section we provide detailed guidance for implementing Gibbs variable selection using BUGS software. The section is further divided into four sub-sections involving the definition of the model likelihood $f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$, the specification of the prior distributions $f(\boldsymbol{\beta}|\boldsymbol{\gamma})$ and $f(\boldsymbol{\gamma})$, and, finally, the direct calculation of posterior model probabilities using BUGS.

3.1 Definition of the Likelihood

The linear predictor of type (1) used in Gibbs variable selection and Kuo and Mallick sampler can be easily incorporated in BUGS using the following code

```
for (i in 1:N) { for(j in 1:p) {z[i,j]<-x[i,j]*b[j]*g[j]}}
for (i in 1:N) {
          eta[i] <-sum(z[i,]) ;
          y[i] ~ distribution(parameter1, parameter2) }
```

where

- `N` denotes the sample size,
- `p` the number of total variables under consideration,
- `x[i,j]` is the i, j component of the data or design matrix $\boldsymbol{X}$,
- `y[i]` is the i element of the response vector $\boldsymbol{y}$,
- `b[j]` is the j element of the parameter vector $\boldsymbol{\beta}$,
- `g[j]` is the inclusion indicator for j element of $\boldsymbol{\gamma}$,
- `z[i,j]` is the i, j component of a matrix which is used to simplify calculations,
- `eta[i]` is the i element of linear predictor vector $\boldsymbol{\eta}$ and should be substituted by the corresponding link function, for example `logit(p[i])` in binomial logistic regression,
- `distribution` should be substituted by the appropriate BUGS command for the distribution that the user prefers (for example `dnorm` for normal distribution),
- `parameter1,parameter2` should be substituted according to distribution chosen, for example for the normal distribution with mean μ_i and variance τ^{-1} we may use `mu[i], tau`.

Alternatively, if we prefer to use SSVS as defined by George and McCulloch (1993) should change the first line of the above code to

```
for (i in 1:N) { for (j in 1:p) { z[i,j]<-x[i,j]*b[j]}}.
```

For the usual normal, binomial and Poisson models the model formulations are given by the following lines of BUGS code

Normal: `for (i in 1:N) { mu[i] <- sum(z[i,]) ;`
```
                         y[i] ~ dnorm (mu[i], tau)  }
```

where `mu[i]` is the expected value for the ith observation and `tau` is the precision of the regression model.

Poisson: `for (i in 1:N) { log(lambda[i]) <- sum(z[i,]);`
```
                          y[i] ~ dpois(lambda[i])}
```

where `lambda[i]` is the Poisson mean for the ith observation.

Binomial: `for (i in 1:N) { logit(p[i]) <- sum(z[i,]);`
```
                           y[i] ~ dbin(p[i], n[i])}
```

where `p[i]` is the probability of success and `n[i]` is the total number of Bernoulli trials for the ith binomial experiment. Alternative link functions maybe used by substituting `logit(p[i])` by `probit(p[i])` or `cloglog(p[i])` for $\Phi^{-1}(p)$ and $log(-log(1-p))$; where Φ is the normal cumulative distribution function.

3.2 Specification of Prior Distribution of Parameter Vector

In situations where we use independent priors similar to (6) and each covariate parameter vector is univariate, the specification of the prior is straightforward. Our prior is a mixture of independent normal distributions

$$\beta_j \sim (1-\gamma_j)N(\bar{\mu}_j, S_j) + \gamma_j N(0, \Sigma_j), \quad j = 1, 2, \ldots, p \tag{9}$$

where $\bar{\mu}_j$, S_j are the mean and variance respectively used in the corresponding pseudoprior distributions and Σ_j is the prior variance, when the j term is included in the model. In order to use (9) in BUGS we write

- `b[j]~dnorm( bpriorm[j], tprior[j])` denoting $\beta_j \sim N(m_j, \tau_j^{-1})$,
- `bpriorm[j] <- (1-g[j])*mean[j]` denoting $m_j = (1-\gamma_j)\bar{\mu}_j$,
- `tprior[j] <- g[j]*t[j]+(1-g[j])*pow(se[j],-2)` denoting $\tau_j = (1-\gamma_j)S_j^{-1} + \gamma_j \Sigma_j^{-1}$,

for $j = 1, 2, \ldots, p$; where m_j and τ_j are the prior mean and precision for β_j depending on γ_j and `t[j], se[j], mean[j], bpriorm[j], tprior[j]` are the BUGS variables for Σ_j^{-1}, $\sqrt{S_j}$, $\bar{\mu}_j$, m_j and τ_j, respectively.

When we have multivariate $\boldsymbol{\beta}_j$ then the vector $\boldsymbol{\beta}$ has greater dimensionality than $\boldsymbol{\gamma}$. In these situations we denote by p and $d(>p)$ the dimensions of $\boldsymbol{\gamma}$ and the full parameter vector $\boldsymbol{\beta}$, respectively. Therefore, we need one variable to facilitate the association between these two vectors. This vector is denoted by the BUGS variable `pos`. The `pos` vector, which has dimension equal to the dimension of $\boldsymbol{\beta}$, takes values from $1, 2, ..., p$ and denotes that β_k is related to γ_{pos_k} for $k = 1, 2, ..., d$.

Here we illustrate the use of a mixture of normal prior distributions as in (6). This prior can be expressed as a multivariate normal distribution on the 'full' parameter vector $\boldsymbol{\beta}$. Therefore we write in BUGS

- `b[ ]` $\sim$ `dmnorm( bpriorm[ ], Tau[,])` denoting $\boldsymbol{\beta} \sim N_d(\boldsymbol{m}, \boldsymbol{T}^{-1})$,
- `bpriorm[k]`$<-$`(1-g[pos[k]])*mean[k]` denoting $m_k = (1-\gamma_{pos_k})\bar{\mu}_k$,
- `Tau[k,l]` $<-$ `(1-g[pos[k]]*g[pos[l]])*t[k,l]+` `+g[pos[k]]*gamma[pos[l]]*equals(k,l)*pow(se[k],-2)` denoting that

$$T_{kl} = \begin{cases} [\boldsymbol{\Sigma}^{-1}]_{kl} & \text{when } \gamma_{pos_k} = \gamma_{pos_l} = 1 \\ se_k^{-2} & \text{when } k = l \ \& \ \gamma_{pos_k} = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } k, l = 1, 2, \ldots, d;$$

where N_d is the d-dimensional normal distribution; $\mathbf{m}^T = (m_1, m_2, \ldots, m_d)$ and $\mathbf{T}$ are the prior mean vector and precision matrix depending on $\boldsymbol{\gamma}$; $\bar{\mu}_k$ is the corresponding pilot run estimate for k element of model parameter vector $\boldsymbol{\beta}$; $\boldsymbol{\Sigma}$ is the constructed prior covariance matrix for the whole parameter vector $\boldsymbol{\beta}$ when we use for each $\boldsymbol{\beta}_j$ the multivariate extension of prior distribution (9); T_{kl} and $[\boldsymbol{\Sigma}^{-1}]_{kl}$ is the k row and l column elements of $\boldsymbol{T}$ and $\boldsymbol{\Sigma}^{-1}$ matrices respectively; and `Tau[,]`, `t[,]` are the BUGS matrices for $\boldsymbol{T}$ and $\boldsymbol{\Sigma}^{-1}$, respectively. For application of the above see example 1.

SSVS and Kuo Mallick sampler can be easily applied by slightly changing the above code. In SSVS the prior (9) is used with $\bar{\mu}_j = 0$ and $S_j = \Sigma_j / k_j^2$, where k_j^2 should be large enough in order that β_j will be close to zero when $\gamma_j = 0$. For selection of the prior parameters in SSVS see semiautomatic prior selection of George and McCulloch (1993). The above restriction can be easily applied in BUGS by

```
bpriorm[j] <- 0
tprior[j] <- t[j]*g[j]+(1-g[j])*t[j]*pow(k[j],2) .
```

Kuo and Mallick sampler uses prior on $\boldsymbol{\beta}$ that does not depend on model indicator $\boldsymbol{\gamma}$. Therefore prior specification is the same as in simple modelling with BUGS.

3.3 Specification of Prior Term Probabilities

In order to apply any variable selection method in BUGS we need to define the prior probabilities $f(\boldsymbol{\gamma})$. When we are vague about models we may set $f(\boldsymbol{\gamma}) =$

$1/|\mathcal{M}|$, where $|\mathcal{M}|$ is the number of all models under consideration. When the explanatory variables do not involve interactions (e.g. linear regression) then the number of models under consideration is 2^p. In these situations the latent variables γ_j can be treated as $a-priori$ independent and therefore set in BUGS

- `g[j]` $\sim$ `dbern(0.5)` denoting that $\gamma_j \sim Bernoulli(0.5)$.

for all $j = 1, 2, \ldots, p$. This prior results to $f(\boldsymbol{\gamma}) = 2^{-p}$ for all $\boldsymbol{\gamma} \in \{0,1\}^p$. When we are dealing with models using categorical explanatory variables with interaction terms, such as $ANOVA$ or log-linear models, we usually want to restrict attention to hierarchical models. The conditional distributions of $f(\gamma_j|\boldsymbol{\gamma}_{\setminus j})$ need to be specified in such way that $f(\boldsymbol{\gamma}) = |\mathcal{M}|^{-1}$ when $\boldsymbol{\gamma}$ is referring to hierarchical model and $f(\boldsymbol{\gamma}) = 0$ otherwise.

For example, in a two way $ANOVA$ we have three terms under consideration ($p = 3$). All possible models are eight, while the plausible models are only five ($constant, [A], [B], [A][B]$ and $[AB]$). Therefore, we need to have $f(\boldsymbol{\gamma}) = 0.20$ for the above five models and $f(\boldsymbol{\gamma}) = 0$ for the rest. This can be applied by setting in BUGS

- `g[3]` $\sim$ `dbern(0.2)` denoting that $\gamma_{AB} \sim Bernoulli(0.2)$.
- `pi` $<-$ `g[3]+0.5(1-g[3])` denoting that $\pi = \gamma_{AB} + 0.5(1-\gamma_{AB})$.
- `for (i in 1:2)` $\{$ `g[j]` $\sim$ `dbern(pi)` $\}$ denoting that for all $i \in \{A, B\}$, $\gamma_j|\gamma_{AB} \sim Bernoulli(\pi)$.

From the above it is evident that

$$\begin{aligned} f([AB]) &= f(\gamma_{AB}=1)f(\gamma_A=1|\gamma_{AB}=1)f(\gamma_B=1|\gamma_{AB}=1) \\ &= 0.2 \times 1 \times 1 = 0.2 \end{aligned}$$

$$\begin{aligned} f([A][B]) &= f(\gamma_{AB}=0)f(\gamma_A=1|\gamma_{AB}=0)f(\gamma_B=1|\gamma_{AB}=0) \\ &= 0.8 \times 0.5 \times 0.5 = 0.2 \end{aligned}$$

Using similar calculations we can find that $f(\boldsymbol{\gamma}) = 0.2$ for all five models under consideration. For further relevant discussion and application see Chipman (1996). For implementation in BUGS see examples 1 and 3.

3.4 Calculating Model Probabilities in BUGS

In order to directly calculate in BUGS the posterior model probabilities and avoid saving large output we can use matrix type variables with dimension equal to the number of models. Using a simple coding such as $1+\sum_{j=1}^{p} \gamma_j 2^{j-1}$ we transform the vector $\boldsymbol{\gamma}$ in a unique, for each model index (noted by `mdl`) for which `pmdl[mdl]=1` and `pmdl[j]=0` for all $\mathtt{j} \neq \mathtt{mdl}$. The above statements can be written in BUGS with the code

```
for (j in 1:p) { index[j] <- pow(2,j-1) }
mdl <- 1+inprod(g[ ], index[ ])
for (m in 1:mdl) { pmdl[m] <- equals(m,mdl) }
```

Then using the command `stats(pmdl)` in BUGS environment (or cmd file) we can monitor the posterior model probabilities. This is feasible only if the number of models is limited and therefore applicable only in some simple cases.

4 Examples

The implementation of of three illustrated examples are briefly presented. The first example is a $3 \times 2 \times 4$ contingency table used to illustrate how to handle factors with more than two levels. Example 2 provides model selection details in a regression type problem involving many different error distributions while example 3 is a simple logistic regression problem with random effects. In all examples posterior probabilities are presented while the associated BUGS codes are provided in the appendix. Additional details (for example, convergence plots) are omitted since our aim is just to illustrate how to use BUGS for variable selection.

4.1 Example 1: $3 \times 2 \times 4$ Contingency Table

This example is a $3 \times 2 \times 4$ contingency table presented by Knuiman and Speed (1988) where 491 individuals are classified by three categorical variables: obesity (O: low,average,high), hypertension (H: yes,no) and alcohol consumption (A: 1,1–2,3–5,6+ drinks per day); see Table 1.

		Alcohol Intake			
Obesity	**High BP**	0	1-2	3-5	6+
Low	Yes	5	9	8	10
	No	40	36	33	24
Average	Yes	6	9	11	14
	No	33	23	35	30
High	Yes	9	12	19	19
	No	24	25	28	29

Table 1: $3 \times 2 \times 4$ Contingency Table: Knuiman and Speed (1988) Dataset.

The full model is given by

$$n_{ilk} \sim Poisson(\lambda_{ilk}), \quad log(\lambda_{ilk}) = m + o_i + h_l + a_k + oh_{il} + oa_{ik} + ha_{lk} + oha_{ilk},$$

for $i = 1, 2, 3,\ l = 1, 2,\ k = 1, 2, 3, 4$. The above model can be rewritten with likelihood given by (1) where $\boldsymbol{\beta}$ can be divided to $\boldsymbol{\beta}_j$ sub-vectors with $j \in \{\emptyset, O, H, OH, A, OA, HA, OHA\}$; where $\beta_\emptyset = m$, $\boldsymbol{\beta}_O^T = [o_2, o_3]$, $\beta_H = h_2$, $\beta_{OH}^T = [oh_{22}, oh_{32}]$, $\boldsymbol{\beta}_A^T = [a_2, a_3, a_4]$, $\boldsymbol{\beta}_{OA}^T = [oa_{22}, oa_{23}, oa_{32}, oa_{33}]$, $\boldsymbol{\beta}_{HA}^T = [ha_{22}, ha_{23}]$ and $\boldsymbol{\beta}_{OHA}^T = [oha_{222}, oha_{223}, oha_{322}, oha_{323}]$. Each $\boldsymbol{\beta}_j$ is a multivariate vector and therefore each prior distribution involves mixture multivariate normal distributions. We use sum-to-zero constraints and prior variance Σ_j as in Dellaportas and Forster (1999). We restrict attention in hierarchical models including always the main effects since we are mainly interested for relationships between the categorical factors. Under these restrictions the models under consideration are nine and in order to forbid to move to non hierarchical models we use the following priors in BUGS

- `g[8] ~ dbern(0.1111)` for $\gamma_{OHA} \sim Bernoulli(1/9)$.
- `pi <- g[8]+0.5(1-g[8])` for $\pi = \gamma_{OHA} + 0.5(1-\gamma_{OHA})$,
- `for (i in 5:7) { g[j]~dbern(pi) }` for $\gamma_j|\gamma_{OHA} \sim Bernoulli(\pi)$ for all $i \in \{OH, OA, HA\}$,
- `for (j in 1:4) { g[j] ~ dbern(1) }` for $\gamma_j \sim Bernoulli(1)$ for all $i \in \{constant, O, H, A\}$.

These priors result to prior probability for all hierarchical models equal to 1/9 and zero otherwise.

Pseudopriors	k=10		Pilot Run	
Burn-in	1,000	10,000	1,000	10,000
Iterations	1,000	$10 \times 10,000$	1,000	$10 \times 10,000$
Models				
$[O][H][A]$	62.80	68.87	65.20	67.80
$[OH][A]$	36.90	30.53	34.40	31.63
$[O][HA]$	0.20	0.40	0.10	0.43
$[OH][HA]$	0.10	0.20	0.30	0.14
Terms				
$\gamma_{OH} = 1$	37.00	30.63	34.70	31.77
$\gamma_{HA} = 1$	0.30	0.20	0.40	0.57

Table 2: $3 \times 2 \times 4$ Contingency Table: Posterior Model Probabilities Using BUGS.

Results using both pilot run pseudoprior and automatic pseudoprior with $k = 10$ are summarised in Table 2. The data give 'strong' evidence in favour of the model of independence. Model [OH][A], in which obesity and hypertension are depending on each other given the level of alcohol consumption, is the model with the second highest posterior probability. All the other models have probability lower than 1%.

4.2 Example 2: Stacks Dataset

Stacks example is a stack-loss data analysed by Spiegelhalter *et al.* (1996b, page 27) using Gibbs sampling. The dataset features 21 daily responses of stack loss (y) which measures the amount of ammonia escaping with covariates the air flow (x_1), temperature (x_2) and acid concentration (x_3). Spiegelhalter *et al.* (1996b) consider regression models with four different error structures (normal, double exponential, logistic and t_4 distributions). They also consider the cases of ridge and simple independent regression models. We extend their work by applying Gibbs variable selection on all these eight cases.

The full model will be

$$y_i \sim D(\mu_i, \tau), \;\; \mu_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3}, \;\; i = 1, \ldots, 21$$

where $D_i(\mu_i, \tau)$ is the distribution of the errors with mean μ_i and variance τ^{-1} which here is assumed to be normal, or double exponential, or logistic or t_4; where $z_{ij} = (x_{ij} - \bar{x}_j)/sd(x_j)$ are the standardised covariates. The ridge regression assumes a further restriction that the β_j for $j = 1, 2, 3$ are exchangeable (Lindley and Smith, 1972) and therefore we have $\beta_j \sim N(0, \phi^{-1})$. We use 'non-informative' priors with prior precision equal to 10^{-3} for the independent regression and for ϕ in ridge regression we use gamma prior with parameters equal to 10^{-3}. Since we do not have restrictions for the model space we use $\gamma_j \sim Bernoulli(0.5)$ for $j = 1, 2, 3$ which results to prior probability of 1/8 for all possible models. For the pilot run pseudoprior parameters we use the posterior values as given Spiegelhalter *et al.* (1996b).

Tables 3 and 4 provide the results from all eight distinct cases using pilot run pseudopriors. In all cases flow of air (z_1) has posterior probability of inclusion higher than 99%. The temperature (z_2) seems to be also an important term with posterior probability of inclusion varying from 39% to 96%. The last term (z_3) which measures the acid concentration in air has low posterior probabilities of inclusion which are less than 5% for simple independence models and less than 20% for 'ridge' regression models.

4.3 Example 3: Seeds Dataset, Logistic Regression with Random Effects

This example involves the examination of a proportion of seeds that germinated on 21 plates. For these 21 plates we have recorded the seed (bean or cucumber) and the type of root extract. This data set is analysed by Spiegelhalter *et al.* (1996b, page 10) using BUGS; for more details see references there in. The model is a logistic regression with 2 categorical explanatory

	Independence Regression			
Models	Normal	D.Exp.	Logistic	t_4
Constant	0.00	0.00	0.00	0.00
z_1	14.12	58.48	41.19	56.46
z_2	0.56	0.01	0.02	0.00
$z_1 + z_2$	81.25	38.64	55.25	40.46
z_3	0.00	0.00	0.00	0.00
$z_1 + z_3$	0.63	1.75	1.35	1.82
$z_2 + z_3$	0.05	0.00	0.00	0.00
$z_1 + z_2 + z_3$	3.39	1.11	2.18	1.26
Terms				
$\gamma_{z_1} = 1$	99.30	99.98	99.97	100.00
$\gamma_{z_2} = 1$	84.90	39.76	57.45	41.72
$\gamma_{z_3} = 1$	4.30	2.86	3.53	3.08

Table 3: Stacks Dataset: Posterior Model Probabilities in Independence Regression (burn-in 10,000, samples of $10 \times 10,000$, with pilot run pseudopriors).

	Ridge Regression			
Models	Normal	D.Exp.	Logistic	t_4
Constant	0.00	0.00	0.00	0.00
z_1	3.26	22.54	14.42	13.30
z_2	0.05	0.00	0.00	0.00
$z_1 + z_2$	79.79	65.00	73.32	70.92
z_3	0.00	0.00	0.00	0.00
$z_1 + z_3$	0.44	1.74	1.32	1.86
$z_2 + z_3$	0.00	0.00	0.00	0.00
$z_1 + z_2 + z_3$	16.46	10.72	11.01	13.92
Terms				
$\gamma_{z_1} = 1$	100.00	100.00	100.00	100.00
$\gamma_{z_2} = 1$	96.50	75.72	84.33	84.84
$\gamma_{z_3} = 1$	16.10	12.46	12.33	15.78

Table 4: Stacks Dataset: Posterior Model Probabilities in Ridge Regression (burn-in 10,000, samples of $10 \times 10,000$, with pilot run pseudopriors).

variables and random effects. The full model will be written

$$y_{ilk} \sim Bin(n_{ilk}, p_{ilk}),\ log\left(\frac{p_{ilk}}{1-p_{ilk}}\right) = m + a_i + b_l + ab_{il} + w_k,$$

for $i, l = 1, 2$ and $k = 1, \ldots, 21$; where y_{ilk} and n_{ilk} is the number of seeds germinated and total number of seeds respectively for i seed, l type of root extract and k plate; w_k is the random effect for the k plate.

We use sum-to-zero constraints for both fixed and random effects. The prior variance used here for the fixed effects is $\Sigma = 4 \times 2$. This prior is equivalent to the one used by Dellaportas and Forster (1999) for log-linear model selection. The prior for the precision of the random effects is considered to be a gamma distribution with parameters equal to 10^{-3}. The pseudoprior parameters were taken from a pilot chain of the saturated model. The models under consideration are ten. The prior term probabilities for the fixed effects are assigned similarly as in the example for two-way ANOVA models. For the random effects term indicator we have that $\gamma_w \sim Bernoulli(0.5)$.

	Fixed Effects		Random Effects	
Models	k=10	Pilot	k=10	Pilot
$Constant$	0.00	0.00	1.21	0.99
$[A]$	0.00	0.00	0.22	0.07
$[B]$	32.34	32.07	50.61	50.75
$[A][B]$	3.78	3.84	7.24	7.60
$[AB]$	2.80	2.83	1.80	1.85
Total	38.92	38.74	61.08	61.26

Table 5: Seeds Dataset: Posterior Model Probabilities Using BUGS (burn-in 10,000, samples of $10 \times 10,000$).

Table 5 provides the calculated posterior model probabilities. We used both pilot run proposals and automatic pseudoprior with $k = 10$. Both chains gave the same results as expected and the type of root extract (B) is the only factor that influences the proportion of germinated gems. The corresponding models with random and fixed effects have posterior probability equal to 51% and 32%, respectively. The marginal posterior probability of random effects is 61% which is about 56% higher than the posterior probability of fixed effects models.

5 Appendix: BUGS Codes

BUGS code and all associated data files are freely available in electronic form at the internet web site http://www.stat-athens.aueb.gr/~jbn/ or by electronic mail request.

5.1 Example 1

```
model log-linear;
#
#       3x2x4 LOG-LINEAR MODEL SELECTION WITH BUGS (GVS)
#       (c) OCTOBER 1996  ATHENS UNIVERSITY OF
#                         ECONOMICS AND BUSINESS
#       (c) REVISED OCTOBER 1997 AUEB
#
#       WRITTEN BY JOHN NTZOUFRAS UNDER THE SUPERVISION
#               OF P.DELLAPORTAS AND J.J.FORSTER
#
#
const
   terms=8, # number of terms
   N = 24;  # number of Poisson cells
var
        include,    # conditional prior probability for gi
        pmdl[9],    # model indicator vector
        mdl,        # code of model
        b[N],       # model coefficients
        mean[N],    # proposal mean used in pseudoprior
        se[N],      # proposal standard deviation used in
                    # pseudoprior
        bpriorm[N], # prior mean for b depending on g
        Tau[N,N],   # model coefficients precision
        tprior[N,N],# prior value for Tau when all terms
                    # are included in model
        x[N,N],     # design matrix
        z[N,N],     # matrix with z_ij=x_ij b_j g_j, used in
                    # likelihood
        n[N],       # Poisson cells
        pos[N],     # position of each parameter
        lambda[N],  # Poisson mean for each cell
        gtemp[N],   # temporary term indicator vector
        g[terms];   # term indicator vector
data pos,n in "ex2.dat", x in 'ex2des.dat',
     mean, se in 'prop2.dat', tprior in 'cov.dat';
inits in "ex2.in";
{
#
#       Design Matrix can was calculated outside the bugs code
#       (e.g. via S-plus) and then can be used as data to avoid
#       useless intensive calculations. The Design matrix here
#       was calculated as a Kronecker product of stz. design
#       matrices for 3,2 and 4 levels
```

```
#
#
#       associate g[i] with coefficients.
#
        for (i in 1:N) {
                gtemp[i]<-g[pos[i]];
                }
#
#       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]*gtemp[j]
                }
        }
#
#       model configuration
        for (i in 1:N) {
                log(lambda[i])<-sum(z[i,])
                n[i]~dpois(lambda[i]);
              }
#       defining model code
#       0 for independence model [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],
#       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
#       Mixture normal depending on current status of g[i]
#
        for (i in 1:N) { for (j in 1:N) {
#
#              GVS using se,mean from pilot run
#              *******************************
#
               Tau[i,j]<-0+tprior[i,j]*(gtemp[i]*gtemp[j])+
               (1-gtemp[i]*gtemp[j])*equals(i,j)/(se[i]*se[i]);
#
#           Automatic proposal using prior similar to SSVS
#                             with k=10
#           ************************************************
#           Tau[i,j]<-tprior[i,j]*pow(100,1-gtemp[i]*gtemp[j]);
#
#            Kuo and Mallick proposal is independent of g[i]
#                      [tau[i]=1/2   and bpriorm[i]=0]
#            ***********************************************
#
#
#             Tau[i,j]<-tprior[i,j];
```

```
#
        }
#
#               GVS PRIOR M FROM PILOT RUN
#               **************************
                bpriorm[i]<-mean[i]*(1-gtemp[i]);
#
#               PRIOR M FOR THAT DOES NOT DEPEND ON G.
#               *************************************
#                 bpriorm[i]<-0.0;
        }
        b[]~dmnorm(bpriorm[],Tau[,]);
#
#       defining prior information for gi in such way that
#       allow only hierarchical models with equal probability.
#       We also ignore models nested to the model of
#       independence [A][B][C] since we are interested in
#       associations between factors.
        g[8]~dbern(0.1111111);
        include<-(1-g[8])*0.5+g[8]*1.0;
        g[7]~dbern(include);
        g[6]~dbern(include);
        g[5]~dbern(include);
        for (i in 1:4) {
                g[i]~dbern(1.0);
        }
}
```

5.2 Example 2

```
model stacks;
#
#       LINEAR REGRESSION VARIABLE SELECTION WITH BUGS (GVS)
#          BUGS EXAMPLE: STACKS, see BUGS examples vol.1
#
#       (c) OCTOBER 1997 ATHENS UNIVERSITY OF ECONOMICS
#
#       WRITTEN BY JOHN NTZOUFRAS UNDER THE SUPERVISION
#               OF P.DELLAPORTAS AND J.J.FORSTER
#
#
const
   p = 3,          # number of covariates
   N = 21,         # number of observations
   models=8,       # number of models under consideration 2^8
   PI = 3.141593;
var
   x[N,p],         # raw covariates
   z[N,p] ,        # standardised covariates
   Y[N],mu[N],     # data and expectations
   stres[N],       # standardised residuals
   outlier[N],     # indicator if |stan res|  > 2.5
   beta0,beta[p], # standardised intercept, coefficients
```

```
   b0,b[p],       # unstandardised intercept, coefficients
   phi,           # prior precision of standardised coef.
   tau,sigma,d,   # precision, sd and d.f. of t distribution
   g[p],          # variable indicators
   mdl,           # Model index
   pmdl[models],  # Vector with model indicators
   mean[p],se[p], # pseudoprior mean and se error
   bprior[p],     # Conditional to model Prior prior mean
   tprior[p];     # Conditional to model Prior prior precision
data Y,x in  "STACKS.DAT",
# files with proposed values
mean,se in 'pnorm.dat';   # Normal distribution
#mean,se in 'pdexp.dat';  # Double exponential distribution
#mean,se in 'plogist.dat';# Logistic distribution
#mean,se in 'pt4.dat';    # Student(4) distribution
inits in "STACKS.IN";
{
# Standardise x's and coefficients
  for (j in 1:p) {
      b[j] <- beta[j]/sd(x[,j]) ;
      for (i in 1:N) {
          z[i,j] <- (x[i,j] -  mean(x[,j]))/sd(x[,j]) ;
      }
  }
  b0<-beta0-b[1]*mean(x[,1])-b[2]*mean(x[,2])-b[3]*mean(x[,3]);

# Model
  d <- 4;  # degrees of freedom for t
  for (i in 1:N) {
#
#     Normal Distribution
#     -------------------
      Y[i] ~ dnorm(mu[i],tau);
#
#     Double Exponential Distribution
#     -------------------
#     Y[i] ~ ddexp(mu[i],tau);
#
#     Logistic Distribution
#     -------------------
#     Y[i] ~ dlogis(mu[i],tau);
#
#     Student t4 Distribution
#     -------------------
#     Y[i] ~ dt(mu[i],tau,d);
#
      mu[i] <- beta0 + g[1]*beta[1]*z[i,1]+g[2]*beta[2]*z[i,2]
                     + g[3]*beta[3]*z[i,3];
      stres[i] <- (Y[i] - mu[i])/sigma;
#
#     if standardised residual is greater than 2.5 then outlier
      outlier[i]<-step(stres[i] -2.5) + step(-(stres[i]+2.5) );
  }
#
#       Defining Model Code
```

```
        mdl<- 1+g[1]*1+g[2]*2+g[3]*4
#
#       defining vector with model indicators
        for (j in 1:models){
        pmdl[j]<-equals(mdl,j);}
# Priors
  beta0 ~  dnorm(0,.00001);
  for (j in 1:p) {
#
#      ******** GVS PRIORS FOR INDEPENDENCE REGRESSION ********
#
#      GVS priors with proposals from pilot run
#       bprior[j]<-(1-g[j])*mean[j];
#       tprior[j] <-g[j]*0.001+(1-g[j])/(se[j]*se[j]);
#
#      GVS priors with proposals a mixture of Normals(0,c^2t^2)
      bprior[j]<-0.0;
      tprior[j] <-pow(100,1-g[j])*0.001;
#
#      ******** GVS PRIORS FOR RIDGE REGRESSION ********
#
#      GVS priors with proposals from pilot run
#       bprior[j]<-(1-g[j])*mean[j];
#       tprior[j] <-g[j]*phi+(1-g[j])/(se[j]*se[j]);
#
#      GVS priors with proposals a mixture of Normals(0,c^2t^2)
#       bprior[j]<-0.0;
#       tprior[j] <-pow(100,1-g[j])*phi;
      beta[j] ~ dnorm(bprior[j],tprior[j]); # coefs independent

  }
  tau ~ dgamma(1.0E-3,1.0E-3);
#
#  phi ~ dgamma(1.0E-3,1.0E-3);
#
#  when we use pilot run based pseudopriors bugs was unable
#  to select update method. Therefore we use an upper limit
#  which makes bugs work with Metropolis instead Gibbs
#
#  phi ~ dgamma(1.0E-3,1.0E-3)I(0,10000);
# standard deviation of error distribution
  sigma <- sqrt(1/tau);             # normal errors
# sigma <- sqrt(2)/tau;             # double exponential errors
# sigma <- sqrt(pow(PI,2)/3)/tau ;  # logistic errors
# sigma <- sqrt(d/(tau*(d-2)));     # errors of t with d d.f.
#
#
# Priors for variable indicators
  for (j in 1:p) { g[j]~ dbern(0.5);}
}
```

5.3 Example 3

```
model seedszrogvs;
#
#               LOGISTIC REGRESSION VARIABLE AND
#            RANDOM EFFECTS SELECTION WITH BUGS (GVS)
#
#          BUGS EXAMPLE: SEEDS, see BUGS examples vol.1
#
#       (c) OCTOBER 1997 ATHENS UNIVERSITY OF ECONOMICS
#
#       WRITTEN BY JOHN NTZOUFRAS UNDER THE SUPERVISION
#               OF P.DELLAPORTAS AND J.J.FORSTER
#
#

const
   terms=4,  # Number of terms under consideration
   models=16,# number of models under consideration 2^4
   N = 21;   # number of samples
var
   alpha0, alpha1, alpha2, alpha12,  # model coefficients
   tau, sigma,       # variance of random effects (tau=1/sigma)
   x1[N], x2[N],     # Design Column for factor a1 and a2
                     # here we used the STZ constraints
   p[N],             # Success probability for Binomial
   n[N],             # Total number of trials for Binomial
   r[N],             # Binomial data
   b[N],             # Random effects  (standardised)
   c[N],             # Random effects c[i] (unstandardised)
   include,          # conditional model probability for
                     # main effects
   g[terms],         # terms indicator vector
   mdl,              # model index
   pmdl[models],     # model indicator
   mean[terms-1],    # proposal mean
   se[terms-1],      # proposal se
   bprior[terms-1],# prior mean for model coefficients
   tprior[terms-1];# prior precision for model coefficients
data r,n,x1,x2 in "seeds.dat", mean,se in 'prop6.dat';
inits in "seeds.in";

{
   alpha0 ~ dnorm(0.0,1.0E-6);  # intercept

        for (j in 1:(terms-1))  {
#      ******** GVS PRIORS ***********
#
#      GVS priors with proposals from pilot run
       bprior[j]<-(1-g[j])*mean[j];
       tprior[j] <-g[j]/8+(1-g[j])/(se[j]*se[j]);
#
#      GVS priors with proposals a mixture of Normals(0,c^2t^2)
#      bprior[j]<-0.0;
#      tprior[j] <-pow(100,1-g[j])/8;
```

```
      }
#
#
   alpha1 ~  dnorm(bprior[1],tprior[1]);  # seed coeff
   alpha2 ~  dnorm(bprior[2],tprior[2]);  # extract coeff
   alpha12 ~ dnorm(bprior[3],tprior[3]);
   tau ~ dgamma(1.0E-3,1.0E-3);           # 1/sigma^2
   for (i in 1:N) {
      c[i] ~ dnorm(0.0,tau);
      b[i] <- c[i] - mean(c[]);    #  make sure b's add to zero
      logit(p[i]) <-alpha0+g[1]*alpha1*x1[i]+g[2]*alpha2*x2[i]
                          +g[3]*alpha12*x1[i]*x2[i]+g[4]*b[i];
      r[i] ~ dbin(p[i],n[i]);
   }

   sigma <- 1.0/sqrt(tau);
#
#       Defining Model Code
        mdl<- 1+g[1]*1+g[2]*2+g[3]*4+g[4]*8
#
#       defining vector with model indicators
        for (j in 1:models){
        pmdl[j]<-equals(mdl,j);}
# Priors for variable indicators
        g[4]~ dbern(0.50);
        g[3]~ dbern(0.20);
        include<-g[3]+(1-g[3])*0.5
        g[2]~ dbern(include);
        g[1]~ dbern(include);
                    }
```

References

Berger, J.O. & Pericchi, L.R. (1996), 'The Intrinsic Bayes Factor for Model Selection and Prediction', *Journal of the American Statistical Association*, **91**, 109–122.

Carlin, B.P. & Chib, S. (1995). 'Bayesian Model Choice via Markov Chain Monte Carlo Methods', *Journal of the Royal Statistical Society B*, **157**, 473–484.

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

Clyde, M., DeSimone, H. & Parmigiani, G. (1996). 'Prediction via Orthogonalized Model Mixing', *Journal of the American Statistical Association*, **91**, 1197–1208.

Dellaportas, P. & Forster, J.J. (1999). 'Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-linear Models', *Biometrika*, (to appear).

Dellaportas, P., Forster, J.J. & Ntzoufras, I. (1998). 'On Bayesian Model and Variable Selection Using MCMC', *Technical Report* **40**, Department of Statistics, Athens University of Economics and Business, Greece.

Dellaportas, P., Forster, J.J. & Ntzoufras, I. (1999). 'Bayesian Variable Selection Using the Gibbs Sampler', *Generalized Linear Models: A Bayesian Perspective* (D. K. Dey, S. Ghosh, and B. Mallick, eds.). New York: Marcel Dekker, (to appear).

George, E.I. & Foster, D.P. (1997). 'Calibration and Empirical Bayes Variable Selection', *Technical Report*, University of Texas at Austin and University of Pennsylvania, USA.

George, E.I. & McCulloch, R.E. (1993). 'Variable Selection via Gibbs Sampling', *Journal of the American Statistical Association*, **88**, 881–889.

Green, P. (1995). 'Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination', *Biometrika*, **82**, 711–732.

Kuo, L. & Mallick, B. (1998). 'Variable Selection for Regression Models', *Sankhyā B*, **60**, 65–81.

Knuiman, M.W. & Speed, T.P. (1988). 'Incorporating Prior Information Into the Analysis of Contingency Tables', *Biometrics*, **44**, 1061–1071.

Lindley, D.V. & Smith, A.F.M. (1972). 'Bayes Estimates for the Linear Model' (with discussion). *Journal of the Royal Statistical Society B*, **34**, 1–41.

O'Hagan, A. (1995). 'Fractional Bayes Factors for Model Comparison', *Journal of the Royal Statistical Society B*, **57**, 99–138.

Raftery, A.E. (1996). 'Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models', *Biometrika*, **83**, 251–266.

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

Spiegelhalter, D., Thomas, A., Best, N. & Gilks, W. (1996b). *BUGS 0.5: Examples Volume 1*, MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

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