Bayesian Constrained Variable Selection

Alessio Farcomeni
alessio.farcomeni@uniroma1.it
"Sapienza" University of Rome

Abstract

By building on the stochastic search approach (George and McCulloch, 1993) we propose a strategy for performing constrained variable selection. We discuss hierarchical and grouping constraints; and introduce anti-hierarchical constraints, in which the inclusion of a variable forces another to be excluded from the model. We prove consistency results about model receiving maximal posterior probability and about the median model Barbieri and Berger (2004), and discuss extension to generalized linear models. The model can be easily implemented and fit using a standard Gibbs sampler.

Keywords: Constraints, Gibbs sampler, hierarchical models, variable selection

1 Introduction

Consider the task of predicting a dependent variable Y from the values of p predictors $X_1, \ldots, X_p$ through some linear model. In this paper we will refer to the predictor X_j as "variable", irrespectively of it being a function of any other predictors or not. There are many cases where one would select variables in groups or in hierarchy, thus satisfying constraints on the final composition of a regression model. For instance one may like to include an interaction or a transformation of variables only if the main effects are included too *(hierarchical variable selection)*; or include all or none of the dummies in a corner point parameterization of a categorical variable *(grouped variable selection)*. This setting includes multi-factor ANOVA, and additive models

with polynomial or nonparametric input variables, in which each component is a linear combination of basis functions obtained from the original predictor. More specific applications include genetics (inclusion of genes in pathways), spatial statistics (inclusion of all or no direction, see Zhao *et al.* (2006)), and others. Another situation in which a large number of hierarchical constraints appear is hereditary wavelet thresholding (Autin *et al.*, 2004), in which detail coefficients are forced to enter into the model whenever higher level coefficients are not thresholded to zero. We note that in multi-factor ANOVA it may not always be sensible to force a hierarchical structure for the model: there may be factors that have an interaction but no main effect (see for instance (Scheffè, 1963, Chap. 4, pag. 94)). Furhter, in certain cases only partial inclusion of a categorical may be of interest (Meyer and Laud, 2002).

In this paper we introduce and model a third class of constraints not previously considered in the literature, which we call *anti-hierarchical* constraints. We refer to an anti-hierarchical constraint between a variable X_i and X_j if X_j *need not* be included in every model in which X_i is included. Anti-hierarchical constraints may be useful for (i) cost/availability constraints in further analyses: when the selected model is used for future prediction, it may be the case that not all covariates can be simoultaneously measured in future observations (for instance in medical diagnoses, industrial quality control, etc), (ii) drug design and similar settings in which certain ingredients can't be mixed (iii) collinearity problems: if two variables are almost perfectly correlated, then only one of the two must be included in the model, and (iv) following only a specific route of transformations for interpretability reasons: we may want to consider powers of X_i larger *or* smaller than 1 (that is, a square, cube, etc., or a square,cubic,quartic root).

While grouping constraints are easily embedded into stepwise methods, until recently to the best of our knowledge there had not been attempts to develop methods for automatic hierarchical variable selection. Yuan and Lin (2006); Kim *et al.* (2006); Zhao *et al.* (2006) fill this gap using generalizations of Lasso (Tibshirani, 1996), that is relying on the maximization of a penalized likelihood. The method of Yuan and Lin (2006) has been extended to logistic regression by Meier *et al.* (2006). All the previous methods focus on grouped variable selection, even if they can accomodate hierarchical constraints. While painstacking, and performing a simultaneous shrinkage and selection, all of such methods involve maximization of the likelihood over parameter sets that may be non-convex, and may not be easy to implement. Another problem of Lasso-related methods is that they may not be consis-

tent in model choice in certain situations (Meinshausen and Bühlmann, 2006; Zou, 2006). The goal of this paper is to show that constrained model selection can be performed with a very simple strategy in a Bayesian framework, which naturally embeds shrinkage of the estimates, and yields consistent model choice under weak conditions. By building on the stochastic search approach for Bayesian variable selection (George and McCulloch, 1993, 1997) we propose a Bayesian method for performing grouped, hierarchical and anti-hierarchical variable selection. Due to the stochastic nature of the search algorithm, even when p is large the computational requirements are low, since promising models will be soon sampled often (George and McCulloch, 1993); thus allowing the user to specify a pletora of transformations and interactions as candidate for the final regression model. To the best of our knowledge, this is the first attempt to put constrained variable selection in an automatic Bayesian framework. It is worth noting however that this possibility is considered in different works (Lahiri (2001),Barbieri and Berger (2004)), even if the common approach consists in the enumeration of the model space. Further, King and Brooks (2001) propose an automatic reversible jump approach for hierarchical loglinear models and Zhao *et al.* (2006) provide an interesting Bayesian interpretation of their Composite Absolute Penalties method. The rest of the paper is as follows: in Section 2 we illustrate our strategy for constrained selection. In Section 2.5 we discuss model choice, while frequentist consistency results in model choice, valid also in the general unconstrained framework, are proved in Section 3 We illustrate the method using simulation and a data set in Section 4, and suggest an extension to generalized linear models in Section 5. In Sections 5.1 and 5.2 we illustrate two examples on GLM with canonical and non-canonical link functions.

2 A Bayesian model for constrained variable selection

Bayesian model selection dates back at least to Atkinson (1978), and usually involves a simple multivariate normal modelling. There has been a huge amount of work on the subject since then, which we will not attempt to review. We just point the reader to Lahiri (2001), Chipman *et al.* (2001), and references therein.

We will focus in this paper on stochastic search variable selection

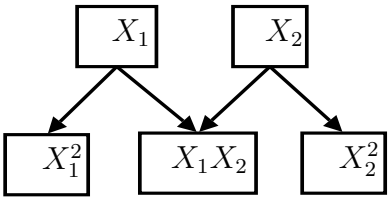

Figure 1: Illustration on a simple model

(George and McCulloch, 1993), where each component of the regression parameter vector β is modelled as a mixture of two centered normal distributions, with different variances. The key feature is the introduction of a binary latent variable γ_j identifying the latent component, and consequently whether the corresponding variable should be included in the final model or not:

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_{0j}^2) + \gamma_j N(0,\tau_{1j}^2); \tag{1}$$

with τ_{1j}^2 slighly larger than τ_{0j}^2. Each model is then identified by a binary vector γ, with prior probability $\pi(\gamma)$, in which the variables corresponding to non-zero components of γ are included and the other are excluded.

We introduce the problem of constrained variable selection with a very simple example.

Example 1. *Suppose we measure two covariates and a continuous response, and consider the possibility to include the square of each measurement and the interaction. The full model is then:*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2.$$

While in certain cases it may be sensible to consider any possible submodel, in many other cases one would like to preserve the hierarchical structure, imposing constraints. The set of constraints can be visualized in Figure 1, where a pointing arrow implies that the parameter cannot be included in the model without the father.

Constrained variable selection could be embedded in this framework by giving zero prior probability to models that do not satisfy the constraints.

This would be anyway a daunting task, that can be avoided with a simple parameterization which we now describe.

We assume the covariates are divided in g *disjoint* groups $G_1, \ldots, G_g$, and define an indicator $\phi_k(j)$ which is 1 if j is member of k-th group. Then, for each variable j we define an indicator function $\delta_j(i)$, $i = 1, \ldots p$, which is 1 if the j-th variable must be included in every model in which i-th is included, and zero if there are no hierarchical constraints. At the same time, we define another indicator $\xi_j(i)$, which is 1 if the j-th variable must be excluded from every model in which i-th is included. Note that ϕ, δ and xi are fixed and specified by the user.

We introduce a second latent indicator η that identifies whether a group is to be included in the final model or not.

Suppose now there is a hierarchical constraint between a "father" variable j_1 in group k_1 and variable j_2 in group k_2. In order to impose this constraint, it is sufficient to set $\gamma_{j_2} = \eta_{k_2}\eta_{k_1}$. In fact, variable j_2 can be included only if *both* groups k_1 and k_2 are included in the model. Since there can be more than one hierarchical constraint, and from different groups, on variable j_2, we make use of indicators δ by setting $\gamma_{j_2} = \eta_{k_2} \prod_{k \neq k_2} \prod_{h \neq j_2} \eta_k^{\delta_h(j_2)\phi_k(h)}$. The last exponent is equal to 1 only if h is a father variable for j_2 and h belongs to the k-th group. A third way to impose the constraints is given by using the other elements of vector γ, and setting: $\gamma_{j_2} = \eta_{k_2} \prod_{j \neq j_2} \gamma_j^{\delta_j(j_2)}$.

Similar approaches can be taken in order to impose grouping and anti-hierarchical constraints.

In the most general case, we link group indicators η to single-variable indicators γ with the following parameterization:

$$\gamma_j(\eta) = \left(\prod_{j \neq i} (1 - \gamma_i(\eta))^{\xi_i(j)} \gamma_i(\eta)^{\delta_i(j)} \right) \prod_{k=1}^{g} \eta_k^{\phi_k(j)}; \tag{2}$$

and for ease of notation we will suppress the dependence of γ_j on η. The indicators ξ, δ and ϕ are specified by the user to identify the constraints and select a subclass of the set of all possible models; and parameterization (2) then directly enforces the constraints. In fact, γ_j can be equal to 1 only if the corresponding η_k is equal to 1 (and thus only together with the members of the same group) and only if all the parent variables are already included in the model. Furthermore, if the i-th variable is included in the model and $\xi_i(j) = 1$, the j-th variable (and consequently its entire hierarchy) is certainly excluded from the model.

2.1 Constraint Specification

A set of constraints formally identifies a subclass of models, and we define:

Definition 1. *A set of constraints given by a choice of ϕ, δ and ξ is called* **compatible** *if for each variable X_j there exist a model in the subclass in which X_j is included.*

Definition 2. *A set of constraints given by a choice of ϕ, δ and ξ is called* **minimal** *if for all i and j any removal of constraints (i.e., changing $\xi_i(j)$, $\delta_i(j)$ or $\phi_k(j)$ from 1 to 0) leads to a different subclass of possible models.*

In practice a set of compatible and minimal constraints allows the user to select each of the variables in at least one model, and each given configuration uniquely identifies a model; which we assume for the rest of the paper.

Since ξ, δ and ϕ are pre-specified, we leave to the user the task of making sure the constraints do not contradict. Constraints contradict for instance if we define an anti-hierarchical constraint between two variables belonging to the same group, thereby excluding the group from the model with probability 1.

It is straightforward to check that problems in compatibility can be experienced *only* when using anti-hierarchical constraints. In all the other cases the set of constraints is compatible.

For instance, if $\xi_i(j)\delta_i(j) = 1$, there is a contradiction and both X_i and X_j will never be selected, thereby having an incompatible set of constraints.

Unfortunately, a universal method to check for compatibility seems hard to develop; and would include checking $\xi_i(j)\delta_i(j) = 0, \forall\, i, j$; $\xi_i(j)\phi_k(i)\phi_k(j) = 0, \forall\, i, j, k$ (i.e., two variables cannot be mutually exclusive and belong to the same group); but also would include checking $\xi_i(j)\delta_i(j_1)\phi_k(j)\phi_k(j_1) = 0$, and so on.

Consequences of not using a compatible set of constraints include almost sure exclusion of certain groups.

Minimality is much less important and aids only computational efficiency. Checking for minimality involves checking for redundancies in constraint specification. If for instance two variables belong to the same group, any hierarchical constraint between them is a redundancy and removing it would lead to the same set of possible models.

Example 2. *In the model of Example 1, there are no grouping and anti-hierarchical constraints, while there are hierarchical constraints. This leads to*

set $\phi_k(j) = 1_{k=j}$*, where* 1_C *is the indicator function of condition* C*,* $\xi_i(j) = 0$ *for all* i, j*;* $\delta_2(1) = \delta_4(3) = \delta_5(1) = \delta_5(3) = 1$ *and* $\delta_i(j) = 0$ *in all other cases.*

2.2 The model

In summary, we propose to fit the the following hierarchical model:

$$\begin{cases} Y \mid \beta, \sigma^2 \sim N(\beta_0 + \sum \beta_k X_k, \sigma^2 I) \\ \sigma^2 \mid \eta \sim IG(\nu_\gamma/2, \nu_\gamma \lambda_\gamma/2) \\ \eta_k \sim \text{Bernoulli} < w_k > \\ \beta \mid \eta \sim N(0, \Gamma R \Gamma). \end{cases} \tag{3}$$

where $\Gamma = \text{diag}(\sqrt{\gamma_j \tau_{1j}^2 + (1-\gamma_j)\tau_{0j}^2})$, and R is a prior correlation matrix. The prior for β leads to a marginal prior as in (1) for each β_j. The only difference with model proposed in George and McCulloch (1993) is that the latent variables γ_j enter into the model as function (2) of group latent variables η_k.

We have few remarks, and illustrate below the main ideas behind the model.

Remark 1. *Grouped and hierarchical constraints are very close in nature. If we allow the groups to overlap, a hierarchical constraint can be specified by inclusion of a group into a larger one. This is the route taken by other works on constrained variable selection. The converse is also true: if we specify that* X_i *is father of* X_j *(*$\delta_i(j) = 1$*) and that* X_j *is father of* X_i *(*$\delta_j(i) = 1$*), then* X_i *and* X_j *belong to the same group, because they can be included in a model only together. We prefer to use separate sets of indicators because it is seen to provide a higher computational stability, and also because we find it more intuitive.*

Remark 2. *The specification of the indicator functions* $\delta_j(i)$ *is very simple, and is itself hierarchical: for instance, for a third order interaction only the second order interactions should be marked as "parents". Marking of the original variables is a redundancy, and can be omitted. Further, anti-hierarchical constraints are reciprocal: if* X_j *need not be included in every model in which* X_i *is included, the converse is also true, and only one direction of the constraint needs to be specified.*

Remark 3. *Note that variables that are not constrained to enter in groups go into a singleton* G_j*. If we only have hierarchical constraints,* $G_j = \{j\}$*,*

$j = 1, \ldots, g$ and $g = p$. In the unconstrained case in which $G_j = \{j\}$, $j = 1, \ldots, p$ and $\delta_j = \xi_j = 0$ for any j, model (3) reduces to the model suggested in George and McCulloch (1993).

Remark 4. *When considering a large number of transformations and interactions, the number of prospective predictors p can get much larger than n. In our Bayesian approach there are no particular problems due to the presence of the prior, making the problem well-posed (see below).*

Remark 5. *In simple cases one can "explode" parameterization (2) and explicitely define each element of the γ vector separately. This is important when sampling with WinBUGS (Lunn* et al.*, 2000) in order to provide a parameterisation with improved orthogonality. As an example, WinBUGS code for the model in Example 1 is given in Appendix B.*

Our setting follows the usual approach of Bayesian variable selection, in which the prior distribution for each β_j coefficient has a spike at zero. In general we set $\tau_{0j}^2 << \tau_{1j}^2$. When $\gamma_j = 0$ and τ_{0j}^2 is small enough, the prior is very concentrated around 0 and values of β_j far from zero receive negligible support. On the other hand, when $\gamma_j = 1$ and τ_{1j}^2 is big enough, a non-zero (posterior) estimate of β_j will probably be included in the final model. The parameter w_k may be interpreted as the statistician's prior probability that variables belonging to group G_k should be included in the final model. We can implicitly penalize the inclusion of variables by setting w_k small enough. The parameter γ_j is a simple parameterization, and is equal almost surely (conditionally on η) to a function of η. Marginally, its prior is a Bernoulli with parameter $\prod_k w_k^{\phi_k(j)} \prod_{i \neq j} \prod_k w_k^{\phi_k(i)\delta_j(i)} (1 - w_k)^{\phi_k(i)\xi_j(i)}$; which can be interpreted as the statistician's prior probability that the predictor X_j is included in the final model, given the constraints. The parameterization and augmentation through the vector η allows to give zero prior probability to models that do not satisfy the constrains in a simple and natural way.

2.3 Sampling from the posterior

The main target of our analysis is the posterior probability of the binary vector γ, $\pi(\gamma \mid Y)$. The vectors γ with higher posterior probability correspond to models receiving higher support by data and prior information.

Sampling from the posterior can be done with the use of classical MCMC methods (Robert and Casella, 1999), and namely by the implementation of a simple Gibbs sampler.

It is straightforward to check that the full conditional for the coefficient vector is

$$\beta \mid Y, X, \sigma^2, \eta \sim N((X'X + D^{-1}R^{-1}D^{-1})^{-1}X'Y, \sigma^2(X'X + D^{-1}R^{-1}D^{-1})^{-1}), \tag{4}$$

where $D = \text{diag}(\sqrt{\gamma_j \tau_{1j}^2 + (1-\gamma_j)\tau_{0j}^2}/\sigma)$. We note that $(X'X + D^{-1}R^{-1}D^{-1})$ is positive definite, hence invertible, for any p.

The full conditional for the variance is instead:

$$\sigma^2 \mid Y, X, \beta, \eta \sim IG(\frac{n+\nu_\gamma}{2}, \frac{\nu_\gamma \lambda_\gamma + |Y - X\beta|^2}{2}) \tag{5}$$

The latent variables η can instead be sampled from:

$$\eta_k | \beta, \sigma^2 \sim \text{Bernoulli} < \frac{w_k a}{w_k a + (1-w_k)b} >, \tag{6}$$

where $a = f(\beta \mid \eta_{-k}, \eta_k = 1) f(\sigma^2 \mid \eta_{-k}, \eta_k = 1)$, $b = f(\beta \mid \eta_{-k}, \eta_k = 0) f(\sigma^2 \mid \eta_{-k}, \eta_k = 0)$ and where η_{-k} stands for the vector η to which the k-th component was removed. It is interesting to note that the full conditional distribution of η_k does not depend on Y, since Y depends on η only through the vector β. If we do not let ν_γ and λ_γ depend on γ, the parameter of the Bernoulli in expression (6) further simplyfies to $\frac{f(\beta|\eta_{-k},\eta_k=1)w_k}{f(\beta|\eta_{-k},\eta_k=1)w_k + f(\beta|\eta_{-k},\eta_k=0)(1-w_k)}$. Finally, γ_j is set equal to the specified function of the vector η.

Even if exploration of all the possible models is not usually feasible, George and McCulloch (1993) note that the fraction of models explored by the MCMC sampling, albeit small, is large enough to narrow down the search for promising variables to be included in the model, and hence for promising models. That is, the models we never see are models we are not interested in.

Note that in general our approach does not involve any transdimensional sampling. There is the possibility to implement more efficient sampling strategies like the one described in Clyde and Parmigiani (1994); Madigan and York (1995); Geweke (1996); George and McCulloch (1997). We note that an additional advantageous possibility is given by the use of adaptive rejection sampling (Gilks and Wild, 1992), which is known to perform well when there is possible multi-modality of the posterior (as in our case for the marginal posteriors of the β parameters).

2.4 Choice of prior parameters

For an informed choice of the prior variance of the coefficients, the same comments in George and McCulloch (1997) apply here: let $\Delta_i = \Delta Y / \Delta X_i$, where ΔY is the size of an insignificant change in Y and ΔX_i the size of a maximum feasible change in X_i. Δ_i is usually referred to as the "threshold of practical significance", since it is believed that whenever $|\beta_i| \leq \Delta_i$ then there is negligible linear relationship between X_i and Y. One can then choose the prior variance so that $\Delta_i^2 = \log(\tau_{1i}^2/\tau_{0i}^2)/(1/\tau_{0i}^2 - 1/\tau_{1i}^2)$, and τ_{1i}^2 is large enough. In general we want to set τ_{0i}^2 small enough so to ensure a posterior estimate close to zero whenever the variable is not relevant in the model, and τ_{1i}^2 big enough so to avoid too much shrinkage towards zero of the posterior estimate if the variable is in fact relevant. The value of τ_{1i}^2 depends then on the order of magnitude of X_i. We have to note however that setting τ_{0i}^2/τ_{1i}^2 too small may slow down the convergence of the MCMC chain, so a long burn-in may be recommended in order to get accurate estimates of β. Standardization can also be used in order to allow for smaller values of τ_{1i}^2. A different possibility is given by setting $\tau_{0i} \cong 0$ and τ_{1i} large (diffuse prior). This is along the lines of the "spike and slab" approach described in Mitchell and Beauchamp (1988), who put a prior probability mass at zero (i.e., $\tau_{0i}^2 = 0$). If τ_{0i}^2 is exactly zero, or too close, then different sampling strategies (for instance, MC^3) may be adopted in order to avoid computational problems and assure convergence of the chain. See for instance Carlin and Chib (1995); Geweke (1996).

If there is no prior information about the probability of inclusion of each group, w_k can be chosen as the indifference probability $w_k = 0.5$. In models with a very large number of predictors, lower values may be more appropriate in order to give higher support to more parsimonius models. For the same reason, another possible choice is anyway to let w_k decrease with the size of the group. If we set equal to p_1 the probability of inclusion of a singleton, the probability of inclusion of group G_k may be set equal to $p_1^{\text{card}(G_k)}$. Note that, due to the model specification, the inclusion of transformations and interactions is directly penalized independently of the choice of w_k. This feature of the model enhances interpretability.

Two common choices are available for the prior correlation matrix. Prior independence is often assumed, in which case R is the identity matrix. Posterior correlations are shrunk towards zero. Another possibility is to have $R \propto (X'X)^{-1}$, in which case posterior correlations are equal to the design

correlation. For further discussion see George and McCulloch (1997); Zellner (1986).

Finally, ν_γ can be as usual interpreted as the prior sample size, and $\nu_\gamma\lambda_\gamma/(\nu_\gamma - 2)$ as a prior estimate for σ^2. One might let ν_γ and λ_γ depend on γ, by having $\nu_\gamma\lambda_\gamma/(\nu_\gamma - 2)$ be decreasing with respect to $\sum\gamma_j$, since it is expected that models in which a larger number of variables is included will be characterized by a smaller residual variance.

2.4.1 Default Priors

Since the main aim of this paper is to cast constrained variable selection in a simple and computationally efficient framework, we proposed the hierarchical model in its simplest form. Such model can be easily generalized to allow for general priors, and additional levels in the hierarchy can be used in order to learn prior parameters.

A particularly relevant setting though is the one given by the use of default priors. The common approach is to combine an improper prior for the intercept and variance of the error term with Zellner's g-prior (Zellner, 1986), thereby having $\pi(\sigma) \propto \sigma^{-1}$ and fixing $R = \sigma^2(hX'X)^{-1}$. This would result in a variable specific g-parameter g_j, set equal to $h/(\gamma_j\tau_{1j}r+(1-\gamma_j)\tau_{0j}^2)$. The tuning parameter h can be chosen so that g_j is equal to one between $1/n$, $1/p^2$, or the smallest between the two. See Fernandez *et al.* (2001) for further discussion. Liang *et al.* (2005) suggest moreover a class of hyperpriors for g which still allow for closed form expressions for the marginal likelihoods. In a similar spirit an hyperprior can be put on w_k as suggested by Ley and Steel (2007).

2.5 Alternative approaches to model choice

Common approaches to model choice rely on the posterior probability of each possible vector γ. Of course, one could simply compute the posterior expected predictive loss corresponding to each model and minimize, but the computational burden (taking the expectation also with respect to future predictors) may be too heavy. From a predictive point of view, it has recently been shown by Barbieri and Berger (2004) that the *median model*, that is, the model in which only variables with posterior probabilities above $\frac{1}{2}$ are included, provides often better predictions than the model with highest posterior probability. Barbieri and Berger (2004) show this result either in

an orthogonal setting, or under conditions that are not very general; but note that the median model is always promising and the only one satisfying optimality results in this sense. They also show that grouped and hierarchical variable selection satisfies the graphical model structure, and thus the median model will always be in the class of possible models. On the other hand, for certain choices of anti-hierarchical constraints, the collection of possible models may violate their condition, and the median model may be outside of the collection. In that case one has to choose the model with highest posterior probability.

Another particularly simple and effective method for model selection is suggested in Madigan and Raftery (1994). Invoking the principle of parsimony and Occam's razor, Madigan and Raftery (1994) suggest discarding all the models for which there is a submodel receiving higher posterior probability, which is particularly appropriate in the setting of hierarchical variable selection.

In the end, assuming the collection satisfy the graphical model structure, this is the novel backward strategy we suggest for model choice:

1. Find the median model. Call it $\mathcal{M}_0$.
2. Set $i := 0$.
3. Consider all the models nested in $\mathcal{M}_i$.
4. If no nested model receives higher posterior support than $\mathcal{M}_i$, the final choice is $\mathcal{M}_i$.
5. If there are nested models receiving higher posterior probability than $\mathcal{M}_i$, consider the one receiving highest posterior probability among them. Call it model $\mathcal{M}_{i+1}$. Set $i := i + 1$.
6. Go to step 3.

This strategy can be applied also in the case of unconstrained variable selection.

Even if we focus here on model choice, we also give some considerations on model averaging (Clyde, 1999; Hoeting *et al.*, 1999). If prediction rather than model choice is the primary goal, it may be more appropriate to use a weighted average of the predictions obtained by conditioning on each possible model; with weight given by the posterior probability of the model.

Constraints may be still useful for at least two reasons: first, a model that is known a-priori not to hold should receive zero posterior probability. Secondly, while many hierarchical constraints would probably be abandoned in model averaging, it may still be appropriate to use anti-hierarchical constraints.

3 Frequentist Properties

We prove in this section certain relevant consistency results for Bayesian variable selection. We point out that these results hold for Bayesian variable selection methods both in the constrained and in the unconstrained case.

Theorem 1. *Assume* $(X'X)/n \to C$, *where* C *is positive definite. We use a short hand notation of* M_0 *for the* γ *vector corresponding to the true model and* $M_{me}|Y$ *for the vector corresponding to the posterior median model. Assume the true and median model are included in the collection of possible models. Fix* $w_k > 0$, $\tau_{0j}^2 < \tau_{1j}^2$ *and*

$$(1 - w_k)\tau_{1j}^2 > w_k \tau_{0j}^2 \tag{7}$$

for all $k = 1, \ldots, g$ *and* $j = 1, \ldots, p$. *Let* ν_γ *and* λ_γ *do not depend on* γ. *Assume also the prior correlation* R *is such that* $\beta_j^* r_{ij}^{-1} \beta_i^* \geq 0$ *for any* i *and* j, *where* r_{ij}^{-1} *is the* ij*-th element of* R^{-1} *and* β^* *is the vector of true parameters. Then,*

$$\lim_{n\to\infty} \Pr(M_{me} = M_0|Y) = 1.$$

If, further, we let $\max_j \tau_{0j} \stackrel{n}{\to} 0$, *then*

$$\lim_{n\to\infty} \Pr(M_0|Y) = 1.$$

Proof. Proof in Appendix A □

The Theorem implies that with minor restrictions on the prior parameters the posterior median model will eventually coincide with the right model, and that the true model will receive posterior probability approaching 1 in the "spike and slab" setting.

The condition (7) of Theorem 1 can be restated in many different convenient ways. A particularly interesting equivalent expression is $0 < w_k < \min_j \frac{\tau_{1j}^2}{\tau_{1j}^2 + \tau_{0j}^2}$, which shows that when $\tau_{0j}^2 << \tau_{1j}^2$ there is very little restriction on the available choices for w_k. Common choices of $w_k \leq 0.5$ satisfy

condition (7) for any $\tau_{1j} > \tau_{0j}$. On the other hand, the condition on R cannot be practically checked, since it depends on the true parameters. It is a sufficient condition requiring coherency between prior beliefs and truth, and could be removed with some restrictions on the magnitude of the prior correlations. Nevertheless, it is straightforward to check that the common choices satisfy the condition: $R = I$ satisfies the condition for any finite n; and $R \propto (X'X)^{-1}$ works asymptotically, which suffices for Theorem 1.

For consistency of the model with highest posterior probability we need to let τ_{0j} decrease to zero. As we pointed out above, there are no particular problems even in setting $\tau_{0j} = 0$ for all n, if one uses an appropriate sampling algorithm.

It is particularly surprising that the results hold without further conditions on X. For instance, orthogonality or other restrictions are needed to prove that the median model is optimal for prediction. The theorem provides in many senses weaker results.

Consistency of the model receiving highest posterior probability has been known for long in the literature. For instance, results dating back at least to Berk (1966), together with Dmochowski (1996), show that under mild conditions common Bayesian methods will choose the right model if it is in the collection, or the closest to the right one in terms of Kullback-Leibler divergence.

To our knowledge consistency results for the median model are instead new also for the case of unconstrained variable selection.

4 Simulations

In order to check the ability of the constrained setting to pick the right model, and prior sensitivity, we simulate different scenarios. First, we generate six covariates $X_1, \ldots, X_6$ from standard normals, and the response from the following model:

$$Y = 1.5X_1 + 2X_2 + X_3 - 1.5X_2X_3 + \varepsilon, \tag{8}$$

where $\varepsilon \sim N(0, 9)$. The sample size is taken to be $n = 250$.

We consider the possibility to include any of the six available covariates, and any of the 15 possible bivariate interactions.

We set $w_k = 0.5$, R to be an identity matrix, and calibrate prior variance parameters to be distant: we set $\tau_0^2 = 0.0625$ and $\tau_1^2 = 1000$.

Scenario	Correct
Scenario 1	0.93
Scenario 2	0.92
Scenario 3	0.96
Scenario 4	0.93
Scenario 5	0.94
Scenario 6	0.66
Scenario 7	0.91

Table 1: Percentage of correct model selection under different scenarios.

We generate the data and use the strategy in Section 2.5 for model choice for $B = 300$ iterations. Percentage of correct model choice is reported in the Scenario 1 row in Table 1.

In Scenario 2 we still simulate from model (8), but also impose an anti-hierarchical constraint between X_4 and X_6. In Scenario 3 we impose two anti-hierarchical constraints, one between X_5 and X_2, and another between X_4 and the interaction between X_2 and X_3. Note that this implies also an anti-hierarchical constraints between X_5 and the interaction.

The third setting is different from the second in that we impose anti-hierarchical constraints between a variable in the true model and one outside. As can be seen from Table 1, this raises the percentage of correct decision, while anti-hierarchical constraints between variables not included in the true model do not seem to have a significant effect.

In Scenario 4 we use the same constraints as in Scenario 1, but set X_4 as $X_4 = X_5 + \varepsilon_2$, where $\varepsilon_2 \sim N(0, 0.01)$. This introduces a strong collinearity in the design matrix, but does not seem to alter the ability of the algorithm to choose the right model.

In Scenario 5 we use the same hierarchical constraints as Scenario 1, and also impose grouping constraints between X_2 and X_3, and between X_5 and X_6.

In Scenario 6 we simulate as in Scenario 1 from model (8), only we consider 25 possible covariates and all their bivariate interactions, ending up with $p = 325 > n = 250$. This is seen to decrease the probability of correct model selection. If nevertheless we use $\tau_1 = 100$ and $\tau_0 = 0.05$ (Scenario 7) the probability of correct model selection raises again (See Table 2 below).

In summary, putting (the right) constraints narrows down the search for

the true model.

Finally, in order to evaluate the effect of the choice of prior parameters, we simulate from Scenario 1 but we try different values for τ_1 and τ_0. Results for $B = 100$ iterations for different combinations of prior parameters are given in Table 2.

We note that a certain degree of dependence on prior choice is well known in Bayesian variable selection, and confirmed by the simulation. Further, τ_1/τ_0 should not be too high.

In conclusion, there *is* an effect of prior parameters, but a reasonable range of choices lead to choose the correct model with high probability.

	$\tau_1 = 5$	$\tau_1 = 10$	$\tau_1 = 100$	$\tau_1 = 200$	$\tau_1 = 500$	$\tau_1 = 1000$
$\tau_0 = 0.1$	1.00	1.00	0.90	0.83	0.75	0.59
$\tau_0 = 0.05$	0.99	1.00	1.00	1.00	1.00	0.98
$\tau_0 = 0.02$	0.98	0.98	1.00	1.00	1.00	0.99
$\tau_0 = 0.005$	0.97	0.99	0.99	0.99	1.00	1.00
$\tau_0 = 0.001$	0.96	0.97	0.98	0.98	0.99	1.00

Table 2: Percentage of correct model selection under Scenario 1 with different priors.

4.1 Birthweight data

We illustrate the model on the birthweight data set from Hosmer and Lemeshow (1989).

We have $n = 189$ observations collected by Baystate Medical Center, Springfield, Massachusetts during 1986. Response is weight at birth, and we have information about mother's age, weight at last menstrual period, race (white, black, other), smoking status during pregnancy, number of previous premature labours, hypertension in the past, uterine irritability and number of physician visits during the first trimester. We consider transformations of numerical variables up to the fourth power, and all possible bivariate interactions; imposing the natural hierarchical constraints. We also have grouping constraints, since we adopt a corner point parameterization for race.

After a burn-in of 50000 iterations we let the Gibbs sampler run for another 50000. The selected model is:

$$\begin{aligned} E[Y|X] &= \texttt{weight} + \texttt{weight}^2 + \texttt{race} + \texttt{uterine irritability} + \\ &\quad + \texttt{hypertens} + \texttt{smoke} + \texttt{hypertens} * \texttt{race}. \end{aligned}$$

In this example, the median model and the model with highest posterior probability coincide.

The same data were analyzed by Yuan and Lin (2006). The selected model substantially coincide with the one suggested by Yuan and Lin (2006), and we also agree in identifying number of visits as the less important covariate (posterior probability: 0.25) and uterine irritability as the most important (posterior probability: 0.92). On the other hand, we include hypertension and weight. Yuan and Lin (2006) considered weight, its square and its cube as a group, concluding it was not important. By imposing hierarchical constraints on the transformation we find that the cube should be very likely excluded, having marginal posterior probability of 0.003. Finally, we speculate the Bayesian hierarchical model selects hypertension because of the presence of an interaction with race.

As seen, the posterior probability of included and excluded variables can be easily evaluated from the MCMC output. We can also record the posterior probability of each sampled model, and plot it in decreasing order. Results are reported in Figure 2, and lead us to conclude that there is a moderate uncertainty in model selection for these data: there does not appear to be a sharp elbow between promising and less promising models, and the number of models sampled at least once is high (872).

We also note that the model that would be chosen by stepwise methods is rather different, and would not respect the hierarchical structure, for instance including the squared weight without the original variable.

5 Extension to Generalized Linear Models

We illustrate now extension to Generalized Linear Models (GLM), see McCullagh and Nelder (1989), also with an example. In GLM the response is assumed to belong to the exponential family:

$$Y \sim \exp\{y\theta - b(\theta)/a(\phi) + c(y, \phi)\},$$

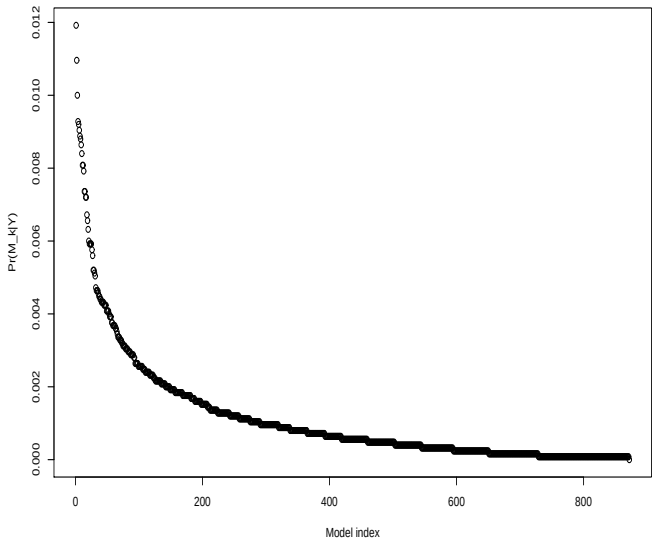

Figure 2: Posterior probability of sampled models (decreasing order) for the Birthweight data set.

with parameters θ and ϕ, and known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$; and a linear relationship is assumed with a function of the expectation of the response:

$$g(E[Y|X]) = \beta_0 + \sum \beta_k X_k,$$

where $g(\cdot)$ is a specified "link function".

It is straightforward to extend our framework to this setting by putting the usual prior structure on the β parameters, and specifying additional priors on nuisance parameters if there are. As before a Gibbs sampler can be set up to simulate from the posteriors. As pointed out by Dellaportas and Smith (1993), also the adaptive rejection method can be used for all canonical link functions (i.e., for $g(\cdot) = b^{'-1}(\cdot)$) and in certain other situations. In cases in which the likelihood function may not be log-concave, one could use the modified version in Gilks *et al.* (1995), which involves a Metropolis step.

5.1 Titanic Data

We illustrate extension to GLM in the context of log-linear models, in which a large number of high-order interactions naturally arise and in which the hierarchical structure shall often be preserved.

Data come from British Board of Trade (1990), who recorded class (1st, 2nd, 3rd or Crew), Sex, Age (adult or child) and survival status for 2201 persons on board of the Titanic, in their investigation of the sinking. Interest in these data stems from the fact that the "women and children first" policy seem not to have been respected for the third class, as reflected by the survival rates.

The class is recoded into three dummy variables (corner point reparameterization), which are grouped, and the other three dummies form three individual groups.

The saturated model includes all the main effects plus interactions up to the fourth order, and can be formulated as:

$$\begin{cases} Y_{ijkh} \sim Poisson(\lambda_{ijkh}) \\ \log(\lambda_{ijkh}) = \beta_0 + \beta_1 class1_i + \ldots + \beta_6 survival_h \\ \qquad +\beta_{14} class1_i * sex_j + \ldots + \beta_{3456} class3_i * sex_j * age_k * survival_h. \end{cases}$$

We want to select a model nested in the saturated model, respecting a hierarchical structure.

We fix $\tau_0 = 0.045$ and $\tau_1 = 10$ and fit the proposed log-linear model on these data forcing a hierarchical structure, the presence of the main effect of survival status; and allowing for interactions up to the fourth order. The posterior median model and model with highest posterior probability coincide, and agree in selecting a log-linear model with main effects, all second-order interactions and all the third-order interactions except one between Sex, Age and Survival.

There is very low uncertainty here in model choice. The selected model has posterior probability 0.51, while the second most likely model only 0.20.

In order to further validate the model we use frequentist measures. The chosen model has likelihood ratio test statistic 1.68, on 4 degrees of freedom (p-value=0.79). The model with better likelihood ratio test statistic with all second-order interactions but only two of the four third-order interactions has likelihood ratio test statistic 21.95, on 7 degrees of freedom (p-value=0.003). Moreover, stepwise methods would lead to select our same model in this case.

5.2 Doctor Visits Data

GLM with noncanonical link functions are often used in practice. We illustrate here an example from the doctor visits data described in Chapter 3 of Cameron and Trivedi (1998). The response is the number of consultations with a doctor or specialist in the previous two weeks, and there are nine predictors: sex, age, age squared, income, health insurance (recoded with three dummy variables), number of illness in the previous two weeks, number of days of reduced activity in the pat two weeks because of illness, general health questionnaire score using Goldberg's method, chronic conditions (recoded with two dummy variables). There are $n = 5190$ observations. The categorical variables make groups, and there is a hierarchical constraint between age and age squared.

The data were analyzed also in Wang and George (2007), who propose the following model:

$$f(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}$$

with (noncanonical) log link function for the mean μ_j. The dispersion parameter α is fixed as its estimated value under the full model.

After MCMC sampling with 20000 iterations and a burn-in of 10000 the median model and the model receiving highest posterior probability coincide; and select sex, age, age squared, illness, days of reduced activity and health score. These results are perfectly in agreement with Wang and George (2007), with the only difference that the model chosen with their preferred method contains the squared age alone in the model, as there is no requirement for a hierarchical structure, while we choose both age and its square because of our constraints.

References

A.C. Atkinson (1978). Posterior probabilities for choosing a regression model. *Biometrika*, **65**, 39–48.

F. Autin, D. Picard, and V. Rivoirard (2004). Maxiset comparisons of procedures, application to choosing priors in a Bayesian nonparametric setting. *Tech. rep.*, Universites de Paris 6 & Paris 7.

M.M. Barbieri and J.O. Berger (2004). Optimal predictive model selection. *Annals of Statistics*, **32**, 870–897.

R. Berk (1966). Limiting behavior of poersterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, **37**, 51–58.

British Board of Trade (1990). *Report on the loss of the Titanic - British Board of Trade Inquiry Report*. Allan Sutton Publishing, Gloucester, UK.

A. Cameron and P.K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge University Press.

B.P. Carlin and S. Chib (1995). Bayesian model choice via Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society (Ser. B)*, **57**, 473–484.

H. Chipman, E.I. George, and R.E. McCulloch (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes - Monograph Series*, **38**, 67–134.

M.A. Clyde (1999). Bayesian model averaging and model search strategies. In: J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., *Bayesian Statistics 6*, 157–185. Oxford Univ. Press.

M.A. Clyde and G. Parmigiani (1994). Bayesian variable selection and prediction with mixtures. *Journal of Biopharmaceutical Statistics*.

P. Dellaportas and A.F.M. Smith (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, **42**, 443–459.

J. Dmochowski (1996). Intrinsic priors via Kullback-Leibler geometry. In: J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., *Bayesian Statistics 5*, 543–549. Oxford Univ. Press.

T.S. Ferguson (1996). *A course in large sample theory*. Chapman & Hall, London.

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

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall.

E.I. George and R.E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

E.I. George and R.E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.

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

W. R. Gilks and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

W.R. Gilks, N.G. Best, and K.K.C. Tan (1995). Adaptive rejection Metropolis sampling within Gibbs sampling (corr: 97v46 p541-542 with Neal, R.M.). *Applied Statistics*, **44**, 455–472.

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volisnky (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382–417.

D.W. Hosmer and S. Lemeshow (1989). *Applied Logistic Regression*. Wiley.

Y. Kim, J. Kim, and Y. Kim (2006). Blockwise sparse regression. *Statistica Sinica*, **16**, 375–390.

R. King and S.P. Brooks (2001). On the Bayesian analysi of population size. *Biometrika*, **88**, 317–336.

P. Lahiri, ed. (2001). *Model Selection*, vol. 38 of *IMS Lecture Notes - Monograph Series*. IMS, Beachwood, OH.

E. Ley and M.F.J. Steel (2007). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Tech. rep.*, University of Warwick, U.K.

F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger (2005). Mixtures of g-priors for Bayesian variable selection. *Tech. Rep. 2005-12*, ISDS Discussion Paper, Duke University.

D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

D. Madigan and A.E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

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

P. McCullagh and J. A. Nelder (1989). *Generalized Linear Models*. Chapman & Hall, CRC, London.

L. Meier, S. van de Geer, and P. Bühlmann (2006). The group Lasso for logistic regression. *Tech. Rep. 131*, Seminar für Statistik.

N. Meinshausen and P. Bühlmann (2006). Variable selection and high-dimensional graphs with the Lasso. *Annals of Statistics*, **34**, 1436–1462.

M.C. Meyer and P.W. Laud (2002). Predictive variable selection in generalized linear models. *Journal of the American Statistical Association*, **97**, 859–871.

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

Christian P. Robert and George Casella (1999). *Monte Carlo statistical methods*. Springer-Verlag Inc.

H. Scheffè (1963). *The analysis of variance*. Wiley, New York.

R. Tibshirani (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society (Ser. B)*, **58**, 267–288.

X. Wang and E.I. George (2007). Adaptive Bayesian criteria in variable selection for generalized linear models. *Statistica Sinica*, **17**, 667–690.

M. Yuan and Y. Lin (2006). Model selection and esimation in regression with grouped variables. *Journal of the Royal Statistical Society (Ser. B)*, **68**, 49–67.

A. Zellner (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: P.K. Goel and A. Zellner, eds., *Bayesian inference and decision techniques: essays in honor of Bruno de Finetti*, 233–243. North-Holland.

P. Zhao, G. Rocha, and B. Yu (2006). Grouped and hierarchical model selection through composite absolute penalties. *Tech. rep.*, University of California, Berkeley.

H. Zou (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

A Proof of Theorem 1

We begin with one preparatory lemma:

Lemma 1. *Suppose $w_k > 0$ for all k, let M_k be an indicator of the k-th model, with $k = 0, \ldots, 2^p - 1$. Let X_{M_k} denote the matrix made of the columns of X corresponding to the variables selected in model M_k, and assume $\frac{1}{n} X'_{M_k} X_{M_k} \to C_{M_k}$, where C_{M_k} is positive definite. Denote also with $\beta^*_{M_k}$ the subvector of β^* with components corresponding to the variables selected in model M_k. Note also that with $\beta \mid M_k$ we refer to the subset of parameters included in model M_k. We have that there exist a product measure $P^\infty_{M_k}$ on $(\mathcal{R}^\infty, \mathcal{B}(\mathcal{R}^\infty))$ such that there exist $\Omega \in \mathcal{B}(\mathcal{R}^\infty))$, of $P^\infty_{M_k}$-probability 1, such that:*

$$\sqrt{n}(\beta \mid M_k, Y - \beta^*_{M_k}) \xrightarrow{d} N(0, \sigma^2 C^{-1}_{M_k}).$$

Further,

$$\sqrt{n}(E[\beta \mid M_k, Y] - \beta^*_{M_k}) \xrightarrow{d} N(0, \sigma^2 C^{-1}_{M_k}),$$

where note that $E[\beta \mid M_k, Y]$ is a random variable as a function of Y.

Proof. It is well known (see for instance Gelman *et al.* (1995)) that the posterior for $\beta \mid M_k, Y$ can be asymptotically approximated by a $N(\beta^*_{M_k}, J(\beta^*_{M_k})^{-1})$, where $J(\beta^*) = (X'_{M_k} X_{M_k})/\sigma^2_{M_k}$ is the Fisher information; provided only that

(i) β is not on the boundary of the parameter space, which is unbounded in our model, and (ii) the likelihood is a continuous function of β. By assumptions we have that $\lim_n J(\beta^*_{M_k})/n = C_{M_k}/\sigma^2$. The second result follows immediately. □

Note that since $\tau_{1j} > 0$ and $w_k > 0$, we essentially are considering the model with all the variables inside. Lemma 1 implies that we have

$$\beta_j|Y \xrightarrow{P} \beta^*. \tag{9}$$

We will repeatedly use the fact that if each element of a finite dimensional vector of random variables converges in probability, then also the vector will converge (see e.g. (Ferguson, 1996, Theorem 6')).

Without loss of generality, let ν_γ and λ_γ not depend on γ; and suppose G_{k_0} is a "father" group, that is, a group of variables for which there are no hierarchical constraints: $\prod\prod \delta_j(i)\phi_{k_0}(i) = 0$. Suppose for simplicity there are no anti-hierarchical constraints in the model.

It is straightforward to check that $\Pr(\eta_{k_0} = 1|Y) = \int \Pr(\eta_{k_0} = 1|\beta)dF(\beta|Y)$. Let the prior correlation R be the identitity matrix. With straightforward computations it can be proved that:

$$\begin{aligned}
\Pr(\eta_{k_0} = 1|\beta) &= \frac{w_{k_0} \prod_{j=1}^{p} \left(1/\tau_{1j} e^{-\frac{1}{2\tau_{1j}^2}\beta_j^2}\right)^{\phi_{k_0}(j)}}{w_{k_0} \prod_{j=1}^{p} \left(1/\tau_{1j} e^{-\frac{1}{2\tau_{1j}^2}\beta_j^2}\right)^{\phi_{k_0}(j)} + (1-w_{k_0}) \prod_{j=1}^{p} \left(1/\tau_{0j} e^{-\frac{1}{2\tau_{0j}^2}\beta_j^2}\right)^{\phi_{k_0}(j)}} \\
&= \frac{1}{1 + \frac{1-w_{k_0}}{w_{k_0}} \prod_{j=1}^{p} \left(\frac{\tau_{1j}}{\tau_{0j}} e^{\frac{\beta_j^2}{2}\frac{(\tau_{0j}^2-\tau_{1j}^2)}{\tau_{0j}^2\tau_{1j}^2}}\right)^{\phi_{k_0}(j)}}.
\end{aligned}$$

If the prior correlation is an arbitrary positive definite matrix, it can be then seen that this only adds an exponential term:

$$\frac{1}{1 + \frac{1-w_{k_0}}{w_{k_0}} \prod_{j=1}^{p} \left(\frac{\tau_{1j}}{\tau_{0j}} e^{\frac{\beta_j^2}{2}\frac{(\tau_{0j}^2-\tau_{1j}^2)}{\tau_{0j}^2\tau_{1j}^2}}\right)^{\phi_{k_0}(j)} \sum_{\eta} \prod_{j:\phi_{k_0}(j)=1} \prod_{i\neq j} e^{\frac{1}{2}\beta_j \frac{(\tau_{0j}-\tau_{1j})}{\tau_{0j}\tau_{1j}} r_{ji}^{-1} \frac{\beta_i}{\gamma_i\tau_{1i}+(1-\gamma_i)\tau_{0i}}} P(\eta|\eta_{k_0}=1)},$$

where r_{ji}^{-1} is the ji-th element of R^{-1}, and we average over all the possible allowed configurations for η (recall that γ_i is function of the vector η).

Suppose now that the k_0-th group is not to be included in the true model. This implies that $\beta_j^* = 0$ for all variables belonging to group G_{k_0}. We need to prove that $\Pr(\eta_{k_0} = 1 \mid Y) < 1/2$ asymptotically. By (9), $\beta_j \xrightarrow{P} 0$ for all j such that $\phi_{k_0}(j) = 1$. It is then straightforward to see that $\Pr(\eta_{k_0} = 1|Y)$ converges to:

$$\frac{1}{1 + \frac{1-w_{k_0}}{w_{k_0}} \prod\limits_{j:\phi_{k_0}(j)=1} \left(\frac{\tau_{1j}}{\tau_{0j}}\right)}. \tag{10}$$

The parameters are tuned by hypothesis so that the previous expression is below 1/2.

If the group corresponding to η_{k_0} must be included in the true model, then $\beta_j^* \neq 0$ for at least one variable belonging to group G_{k_0}. Let j_0 be one of the indices for which $\beta_j^* \neq 0$. We need to prove that $\Pr(\eta_{k_0} = 1 \mid Y) > 1/2$ asymptotically.

Define

$$\theta_j = \sum_{\eta} \prod_{i \neq j} e^{\frac{1}{2}\beta_j^* \frac{(\tau_{0j}-\tau_{1j})}{\tau_{0j}\tau_{1j}} r_{ji}^{-1} \frac{\beta_i^*}{\gamma_i \tau_{1i} + (1-\gamma_i)\tau_{0i}}} P(\eta|\eta_{k_0} = 1).$$

Since by hypothesis $\beta_j^* r_{ij}^{-1} \beta_i^* \geq 0$ for every i and j, it is seen that $\theta_j \leq 1$ for every j. We then have:

$$\begin{aligned}
\Pr(\eta_{k_0} = 1 \mid Y) \;\rightarrow\; & \frac{1}{1 + \frac{1-w_{k_0}}{w_{k_0}} \prod\limits_{j:\phi_{k_0}(j)=1} \left(\frac{\tau_{1j}}{\tau_{0j}}\right) \prod\limits_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} e^{\frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2}} \theta_j} \\
\geq\; & \frac{1}{1 + \frac{1-w_{k_0}}{w_{k_0}} \prod\limits_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} \left(\frac{\tau_{1j}}{\tau_{0j}}\right) e^{\frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2}}} \\
\geq\; & \frac{1}{1 + \frac{(1-w_{k_0})\tau_{1j_0}}{w_{k_0}\tau_{0j_0}} \prod\limits_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} e^{\frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2}}},
\end{aligned} \tag{11}$$

where at the first step we used the fact that $\theta_j \leq 1$ and then repeatedly used the condition that $\tau_{0j}^2 \leq \tau_{1j}^2$.

Last expression is not smaller than $1/2$ if and only if

$$\frac{(1-w_{k_0})\tau_{1j_0}}{w_{k_0}\tau_{0j_0}} \prod_{j:\phi_{k_0}(j)=1\cap\beta_j^*\neq 0} e^{\frac{(\beta_j^*)^2}{2}\frac{(\tau_{0j}^2-\tau_{1j}^2)}{\tau_{0j}^2\tau_{1j}^2}} < 1/2 \Leftarrow$$

$$\prod_{j:\phi_{k_0}(j)=1\cap\beta_j^*\neq 0} e^{\frac{(\beta_j^*)^2}{2}\frac{(\tau_{0j}^2-\tau_{1j}^2)}{\tau_{0j}^2\tau_{1j}^2}} < 1 \Leftrightarrow$$

$$\sum_{j:\phi_{k_0}(j)=1\cap\beta_j^*\neq 0} \frac{(\beta_j^*)^2}{2}\frac{(\tau_{0j}^2-\tau_{1j}^2)}{\tau_{0j}^2\tau_{1j}^2} \leq 0 \Leftarrow$$

$$\tau_{0j}^2 \leq \tau_{1j}^2,$$

which is true by hypothesis. Note that at the second step we used condition (7).

The same results directly follow also for groups for which there are variables with hierarchy constraints, since the probability of selecting the group will only have additional multiplicative terms depending on the probabiliy of selecting the groups in which there are the father variables. Similarly, anti-hierarchical constraints will only lead to the presence of additional multiplicative terms depending on the probability of not selecting groups in which there are the corresponding variables.

The thesis follows: $\Pr(M_{me} = M_0|Y) \to 1$.

To prove the second part, note that $\tau_{0j} \cong 0$ implies that whenever $\eta_k = 0$ all the corresponding βs are zero with probability approaching 1.

By looking at expressions (10) and (11), it is straightforward to check that $\Pr(\eta_k = 1|Y)$ converges to 1 if the k-th group shall be included in the final model and to 0 otherwise since τ_{0j} is infinitesimal.

Without loss of generality assume the true model M_0 is identified by the inclusion in the model of the first k_0 groups and exclusion of the remaining groups.

$$\lim_n \Pr(M_0|Y) = \lim_n \Pr(\cap_{k=1}^{k_0}\eta_k = 1 \cap_{k=k_0+1}^{g} \eta_k = 0|Y), \tag{12}$$

and the right hand side converges to 1 because each element of the vector converges.

B Sample WinBUGS Code for Example 1

```
model
   {

     for(j in 1:N) {
       Y[j] ~ dnorm(mean[j] , S);
       mean[j] <- beta0 + beta1*X[j,1]+ beta2*X[j,1]*X[j,1]
               + beta3*X[j,2] + beta4*X[j,2]*X[j,2] + beta5*X[j,1]*X[j,2];
       }

       beta0  ~ dnorm(0, tau1);

       p1 <- (1-eta1)*tau0+eta1*tau1;
       eta1 ~ dbern( w1);
       beta1  ~ dnorm(0, p1);

       p2 <- (1-gamma2)*tau0+gamma2*tau1;
       gamma2 <- eta1*eta2;
       eta2 ~ dbern( w2);
       beta2  ~ dnorm(0, p2);

       p3 <- (1-eta3)*tau0+eta3*tau1;
       eta3 ~ dbern( w3);
       beta3  ~ dnorm(0, p3);

       p4 <- (1-gamma4)*tau0+gamma4*tau1;
       gamma4 <- eta3*eta4;
       eta4 ~ dbern( w4);
       beta4  ~ dnorm(0, p4);

       p5 <- (1-gamma5)*tau0+gamma5*tau1;
       gamma5 <- eta1*eta3*eta5;
       eta5 ~ dbern( w5);
       beta5 ~ dnorm(0, p5);

       S ~ dchisqr( ds );
   }
```