Bayesian Constrained Variable Selection

Alessio Farcomeni

Sapienza - University of Rome

Supplementary Material

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.

S1 Implementation of the method

In this section we will briefly sketch some guidelines for implementation of the method, summarizing material from Section 2 of the main paper.

We first give a general method for fixing the constraints. We note that in the present formulation constraints specification depends on the column ordering of the variables in the data matrix.

Without loss of generality suppose the groups are ordered so that the first group runs from the first to the j_1-th column, the second from the $j_1 + 1$-th to the j_2-th and so on. Let $j_0 = 1$. The grouping constraints can then simply be set by fixing $\eta_i(j_i) = 1$ for $j = j_{i-1}, \ldots, j_i$ and for $i = 1, \ldots, g$; and $\eta_{\mathbf{i}} = 0$ otherwise.

A general method for fixing hierarchical constraints consists in forming an r_H x 2 matrix H in which in the first column one specifies the position of the father in the data matrix, and in the second the position of the son. Then for $r = 1, \ldots, r_H$ $\gamma_{h_{r1}}(h_{r2}) = 1$, and zero otherwise. The anti-hierarchical constraints can be fixed similarly.

Different approaches can be used depending on how data are organized.

We now focus on the simplest Gibbs sampling strategy that can be adopted for model fitting. Suppose prior parameters have been chosen, together with starting values for the MCMC sampler.

The general iteration of the sampler is as follows:

1. For $j = 1, \ldots, p$ set $\gamma_j = \prod_{k=1}^{g} \eta_k^{\phi_k(j)}$.

2. Let $\gamma_{old} := \gamma$.

$$\text{For } j \in \{j : \prod_{k=1}^{g} \eta_k^{\phi_k(j)} = 1\} \text{ set } \gamma_j := \left(\prod_{j \neq i} \prod_{j \neq i} (1 - \gamma_i)^{\xi_i(j)} \gamma_i^{\delta_i(j)} \right).$$

3. If γ_{old} is equal to γ, go to Step 4, otherwise go to Step 2.

4. Sample β from its full conditional

$$\beta \mid Y, X, \sigma^2, \eta \sim N((X'X + D^{-1}R^{-1}D^{-1})^{-1}X'Y, \sigma^2(X'X + D^{-1}R^{-1}D^{-1})^{-1}), \tag{S1.1}$$

where $D = \text{diag}(\sqrt{\gamma_j \tau_{1j}^2 + (1 - \gamma_j)\tau_{0j}^2}/\sigma)$.

5. Sample σ^2 from its full conditional

$$\sigma^2 \mid Y, X, \beta, \eta \sim IG(\frac{n + \nu_\gamma}{2}, \frac{\nu_\gamma \lambda_\gamma + |Y - X\beta|^2}{2}) \tag{S1.2}$$

6. For $k = 1, \ldots, g$ sample

$$\eta_k \mid \beta, \sigma^2 \sim \text{Bernoulli}(\frac{w_k a}{w_k a + (1 - w_k) b}), \tag{S1.3}$$

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.

S2 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 is 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 group made up of a single variable, 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 approach 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.

We stress that a careful tuning, as in all stochastic variable selection methods, is very important for a quick convergence of the MCMC sampler.

S2.1 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. (2008) 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 (2009).

S3 Real Data Examples

S3.1 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:

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

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 the 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 best 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.

S3.2 Spam Data

In order to illustrate the potentiality of our method with many variables and many constraints, we show application to Spam identification with the the Spambase data set. We have $n = 4601$ emails, 39% of which are spam, and $p = 57$ variables, data and a full

description are available on the UCI Machine Learning Repository (`http://www.ics.uci.edu/ mlearn/MLRepository.html`). The binary response records whether an email is spam or not, and the explanatory variables record frequency of occurrence of certain *flag* words and of special characters. A complete list is in Table 1.

It is natural to expect high-order interactions between the occurrence of certain words in this data set. We proceed by randomly splitting the data set in a training set of 3221 observations and a test set of the remaining 1380.

We consider the possibility of including any of the 54 standardized explanatory variables, transformations up to the power of four, all two way interactions, all two way interactions between the squared variables, and all two way interactions between the squared and the original variables. The resulting number of variables is 5940.

As usual we will not include a power of any order without all the preceeding, and an interaction without the original variables. There is a very large number of complex constraints, which can not be easily exploded. Nevertheless, the hierarchical constraints are easily specified by forming a 5940 by 5940 binary matrix δ containg the $\delta_i(j)$ parameters. For $i = 55, \ldots, 108$ we set $\delta_i(i - 54) = 1$ in order to impose the constraints between the original and the squared variables, for $i = 109, \ldots, 162$ we set $\delta_i(i - 54) = 1$ in order to impose the constraints between the cubes and the squares (and, automatically, between the cubes and the originals); and so on.

In order to penalize more complex models, we set $w_k = 0.05$. The prior variances are set as $\tau_1 = 5$ and $\tau_0 = 0.3$. As noted also in Gustafson and Lefebvre (2008), with such a large model space a true scaling of the posterior is unlikely to occur with few thousands of iterations of the Gibbs sampler. Common sampling schemes are not feasible in the presence of a very large model space. In order to sample from the model we adapt here a parallel search strategy developed in a different context by Corander et al. (2006).

A number m of parallel Gibbs samplers are run, each of which is started from a different point of the model space. In this way, several local neighborhoods are explored. The parallel chains interact so that none is trapped around a local maximum of the likelihood. More in detail, define a binary process Z_t, where t indicizes the chain iterations. Whenever $Z_t = 0$, the chains run in parallel. When $Z_t = 1$, the chains interact in that each is allowed to jump to the state of one of the other chains according to the scheme:

$$\pi(\theta_{t+1j} = \theta_{ti}) = \frac{\pi(\theta_{ti}) f(Y \mid \theta_{ti}, X)}{\sum_{i=1}^{m} \pi(\theta_{ti}) f(Y \mid \theta_{ti}, X)},$$

where θ_{ti} is a short hand notation for the vector of parameters sampled at the t-th iteration of the i-th chain, and $f(Y \mid \theta_{ti}, X)$ is the likelihood. In this way, chains trapped in local modes are allowed from time to time to jump to regions of the model space with higher posterior probability. The process is such that $P(Z_t = 1) = (q \log t)^{-1}$ for $t \geq 1$ and $P(Z_0 = 1) = 0$. Corander et al. (2006) give more details about the advantages of the parallel search strategy, and about consistent estimation of the posterior probabilities after implementation of the parallel sampling scheme.

For these data we use $m = 20$ chains, $q = 5$. Among the starting solutions for the regression coefficients we include the maximum likelihood estimates for the model with only the untransformed variables, and the estimates obtained for the full model with the elastic net penalized likelihood of Zou and Hastie (2005). Among the initial values for η we use the full and empty models, and the model selected by the elastic net.

The median model is made of 360 variables, 43 of which are the original untransformed, together with 36 squares, 2 cubes and no fourth powers. All the remaining selected covariates are interactions. A summary (without the interactions) is in Table 1.

There is very little uncertainty for the chosen model. This can be appreciated for instance from a barplot of $\pi(\gamma_j \mid Y)$ in Figure 1, which shows that the posterior inclusion probabilities are mostly close to zero or to 1 (x-axis: variable index, y-axis: corresponding $\pi(\gamma_j \mid Y)$; the indexes are sorted in decreasing order of $\pi(\gamma_j \mid Y)$). There are only 78 $\pi(\gamma_j \mid Y)$ between 0.3 and 0.7. For instance, there is some uncertainty related to the words "make", which is included only with inclusion probability 0.63 and "mail", which is excluded but has inclusion probability 0.46.

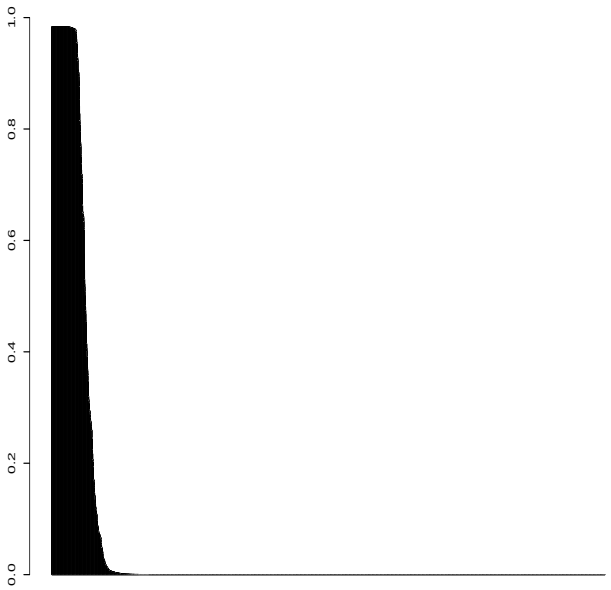

Figure 1: Posterior probability of γ indicators, SPAM data.

The selected model is interpretable. For instance, the frequency of occurrence of the semicolon is discarded, while the frequency of occurrence of the name of the owner of the mailbox is included with very high probability. Of course, emails addressing the receiver

by name are much less likely to be spam. Other words with markedly negative log-odds are "hp" and "meeting". With very high probability we also include the frequency of words "order", "technology", and "000", which may be common words in spam emails.

There are many squared variables included in the model. Some of them have a log-odds coefficient with different sign than the original variable, others have a regression coefficient concordant with the coefficient of the original variable. Parameters of the first kind include "remove", "technology", "business" and the frequency of the character $. Parameters of the second include "hp", "lab", "000". This facts can be easily interpreted. For instance, a larger number of occurrences of the word "hp" should raise the probability of the email not being spam more than linearly. On the other hand, words like "technology" and "business" are often used with high frequency and a larger number of occurrences of those words should not have an high impact on the prediction of the response.

Many words interact with the name of the owner and the word "hp". Other interactions form a part of a sentence, like: "our" and "meeting". When the two words are used together, it is less likely for the mail to be spam and the negative coefficient for the interaction catches this feature. Another interesting interaction is between "hp" and "technology". In fact, a large frequency of the word technology may indicate a spam, but if the word is used in conjuction with "hp", the company of the owner of the mailbox, it is much less likely for the email to be spam. This is reflected on the negative coefficient of the interaction, whereas "technology" has got positive coefficient. Not surprisingly, even if the frequency of occurrence of "000" is very important in the model, there are only 6 interactions with this variable.

Finally, we use the selected model for prediction on the test set. The results are shown in Table 2, where 1-nn and 3-nn stand for the k-nearest neighbours method of Cover and Hart (1967), with respectively $k = 1$ and $k = 3$. Note that no variable selection is available for the k-nn methods. For all the other methods, a (constrained or unconstrained) SSVS is used.

The prediction performance of the constrained model is good, even if not markedly better than the other classification methods. There is a small advantage of using transformed variables and interactions (the proportion of correctly classified emails raises from 90.4% to 91.6%). If the transformations are used without constraints, the prediction performance is not as good likely due to over adaptation to the training set. Moreover the resulting model is not easily interpretable and not as parsimonious, since it uses 711 variables. The logistic model with hierarchical contraints can be used not only for prediction but also for explaining why an email is spam.

S3.3 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 past 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. We use a corner point reparameterization for the categorical variables, and put grouping constraints on the resulting dummies. We also force a hierarchical constraint between age and age squared.

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

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

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.

S4 Proof of Theorem 1

We begin with one preparatory lemma:

Lemma 1. *Let M_k be an indicator for the k-th model in the collection of possible models. Let X_{M_k} denote the matrix made of the columns of X corresponding to the variables used by 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 vector of true parameters for model M_k. Note also that with β_{M_k} we refer to the subset of parameters included in model M_k. By conditioning on M_k we mean that model M_k is deemed to be the true model. We have that*

$$\Pr(\sqrt{n}(\beta_{M_k} - \beta^*_{M_k}) \leq t \mid M_k, Y) \to \Pr(N(0, \sigma^2 C^{-1}_{M_k}) \leq t)$$

for any $t \in \mathcal{R}$, that is β_{M_k} converges in distribution to the true vector of parameters. Further,

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

where note that $E[\beta_{M_k} \mid M_k, Y]$ is a random variable as a function of Y, and $\xrightarrow{d}$ denotes convergence in distribution.

Proof. It is well known (see for instance Gelman et al. (1995)) that if M_k is assumed as the true model, the posterior $\pi(\beta_{M_k} \mid M_k, Y)$ 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. The only conditions needed are that $\beta^*_{M_k}$ is not on the boundary of the parameter space, and that the likelihood is a continuous function of β. The two conditions are met by the proposed model. The first result follows since 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

$$\beta_j \mid Y \xrightarrow{P} \beta^*. \tag{S4.4}$$

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

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

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

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

where r^{-1}_{ji} 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 (S4.4), $\beta_j \xrightarrow{P} 0$ for all j such that $\phi_{k_0}(j) = 1$. It is then straightforward to see that $\Pr(\eta_{k_0} = 1 \mid Y)$ converges to:

$$\frac{1}{1 + \frac{1-w_{k_0}}{w_{k_0}} \prod_{j:\phi_{k_0}(j)=1} \left(\frac{\tau_{1j}}{\tau_{0j}}\right)}. \tag{S4.5}$$

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

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

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:

$$\begin{aligned}
\Pr(\eta_{k_0} = 1 \mid Y) \quad \rightarrow \quad & \frac{1}{1 + \frac{1-w_{k_0}}{w_{k_0}} \prod_{j:\phi_{k_0}(j)=1} \left(\frac{\tau_{1j}}{\tau_{0j}}\right) \prod_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} e^{\frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2}} \theta_j} \\
\geq \quad & \frac{1}{1 + \frac{1-w_{k_0}}{w_{k_0}} \prod_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} \left(\frac{\tau_{1j}}{\tau_{0j}}\right) e^{\frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2}}} \\
\geq \quad & \frac{1}{1 + \frac{(1-w_{k_0})\tau_{1j_0}}{w_{k_0}\tau_{0j_0}} \prod_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} e^{\frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2}}},
\end{aligned} \tag{S4.6}$$

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

$$\begin{aligned}
\frac{(1-w_{k_0})\tau_{1j_0}}{w_{k_0}\tau_{0j_0}} \prod_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} e^{\frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2}} \quad & < \quad 1/2 \Leftarrow \\
\prod_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} e^{\frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2}} \quad & < \quad 1 \Leftrightarrow \\
\sum_{j:\phi_{k_0}(j)=1 \cap \beta_j^* \neq 0} \frac{(\beta_j^*)^2}{2} \frac{(\tau_{0j}^2 - \tau_{1j}^2)}{\tau_{0j}^2 \tau_{1j}^2} \leq 0 \Leftarrow & \\
\tau_{0j}^2 \leq \tau_{1j}^2, &
\end{aligned}$$

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: $\Pr(M_{me} = M_0 \mid Y) \to 1$.

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 (S4.5) and (S4.6), it is straightforward to check that $\Pr(\eta_k = 1 \mid Y)$ 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.

$$\lim_n \Pr(M_0 \mid Y) = \lim_n \Pr(\cap_{k=1}^{k_0} \eta_k = 1 \cap_{k=k_0+1}^{g} \eta_k = 0 \mid Y), \tag{S4.7}$$

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

S5 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 );
}
```

Table 1: Selected model (without list of interactions), SPAM data

	Name	Original	Square	Cube
1	word_freq_make	1	0	0
2	word_freq_address	1	1	0
3	word_freq_all	1	0	0
4	word_freq_3d	1	1	0
5	word_freq_our	1	1	0
6	word_freq_over	1	1	1
7	word_freq_remove	1	1	0
8	word_freq_internet	0	0	0
9	word_freq_order	1	1	0
10	word_freq_mail	0	0	0
11	word_freq_receive	0	0	0
12	word_freq_will	1	1	0
13	word_freq_people	0	0	0
14	word_freq_report	0	0	0
15	word_freq_addresses	1	1	0
16	word_freq_free	1	1	1
17	word_freq_business	1	1	0
18	word_freq_email	0	0	0
19	word_freq_you	0	0	0
20	word_freq_credit	1	0	0
21	word_freq_your	1	1	0
22	word_freq_font	1	1	0
23	word_freq_000	1	1	0
24	word_freq_money	1	1	0
25	word_freq_hp	1	1	0
26	word_freq_hpl	1	1	0
27	word_freq_george	1	1	0
28	word_freq_650	0	0	0
29	word_freq_lab	1	1	0
30	word_freq_labs	1	1	0
31	word_freq_telnet	1	1	0
32	word_freq_857	1	1	0
33	word_freq_data	1	1	0
34	word_freq_415	1	1	0
35	word_freq_85	1	1	0
36	word_freq_technology	1	1	0
37	word_freq_1999	1	1	0
38	word_freq_parts	0	0	0
39	word_freq_pm	1	0	0
40	word_freq_direct	1	0	0
41	word_freq_cs	1	1	0
42	word_freq_meeting	1	1	0
43	word_freq_original	1	1	0
44	word_freq_project	1	1	0
45	word_freq_re	1	1	0
46	word_freq_edu	1	1	0
47	word_freq_table	0	0	0
48	word_freq_conference	1	1	0
49	char_freq_;	0	0	0
50	char_freq_(	1	0	0
51	char_freq_[	1	0	0
52	char_freq_!	1	1	0
53	char_freq_$	1	1	0
54	char_freq_#	1	1	0

Table 2: Prediction on the test set, SPAM data

	Correct	Spam Correct	NonSpam Correct
Logistic model, with constraints	91.6%	94.9%	86.6%
Logistic model, only original variables	90.4%	95.8%	82.2%
Logistic model, without constraints	87.4%	91.5%	84.8%
1-nn	85.4%	88.7%	80.3%
3-nn	85.1%	90.2%	77.3%
1-nn, only original variables	90.0%	91.5%	87.7%
3-nn, only original variables	89.6%	91.8%	86.0%

Bibliography

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

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

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

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

Corander, J., M. Gyllenberg, and T. Koski (2006). Bayesian model learning based on a parallel MCMC strategy. *Statistics and Computing 16*, 355–362.

Cover, T. and P. Hart (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory IT-13*, 21–27.

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

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

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

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

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

Geweke, J. (1996). Variable selection and model comparison in regression. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 5*, pp. 609–620. Oxford Press.

Gustafson, P. and G. Lefebvre (2008). Bayesian multinomial regression with class-specific predictor selection. *Annals of Applied Statistics 2*, 1478–1502.

Ley, E. and M. Steel (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics 24*, 651–674.

Liang, F., R. Paulo, G. Molina, M. Clyde, and J. Berger (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association 103*, 410–423.

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

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

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

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society (Ser. B) 67*, 301–320.