Statistica Sinica **20** (2010), 000-000

BAYESIAN CONSTRAINED VARIABLE SELECTION

Alessio Farcomeni

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 models receiving maximal posterior probability, and about the median model (Barbieri and Berger (2004)), and discuss extensions to generalized linear models.

Key words and phrases: 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 refer to the predictor X_j as a "variable", irrespective of it being a function of any other predictors or not. There are many cases in which one would select variables in groups or in hierarchy, thus satisfying constraints on the final composition of a regression model. Classical constraints include the use of an interaction or a transformation only if the main effects are also included (*hierarchical variable selection*), or the use of all 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. There are specific applications in genetics (inclusion of genes in pathways, epistasis (Cordell (2002))), spatial statistics (inclusion of all or no direction, see Zhao, Rocha and Yu (2009)), and others. Another situation in which a large number of hierarchical constraints appear is hereditary wavelet thresholding (Autin, Picard and Rivoirard (2004)), in which detail coefficients are forced to enter 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)). Further, in certain cases only partial inclusion of a categorical variable may be of interest (Meyer and Laud (2002)).

In this paper, besides the use of hierarchical and grouping constraints, we introduce a third class of *anti-hierarchical* constraints. We refer to an anti-hierarchical constraint between a variable X_i and X_j when X_j *cannot* be included in any model in which X_i is included. Anti-hierarchical constraints may be useful in the following situations: (i) cost/availability reasons (when the selected model is used for prediction, it may be the case that not all covariates can be simultaneously measured in future observations, for instance in medical diagnoses, industrial quality control, etc.); (ii) drug design and similar settings in which certain ingredients cannot be mixed; (iii) collinearity problems associated with the need of avoiding simultaneous use of strongly correlated variables; (iv) for interpretability reasons when some variables are a function of some of the others, for instance in medical research when the MELD score (Cholongitas et al. (2006)) together with its ingredients (patient's creatinine, bilirubin, etc.) are included in the data matrix, but not in the same model; (v) when different transformations are considered, such as allowing for powers of X_j larger *or* smaller than 1.

While the simplest grouping constraints are easily embedded into stepwise methods, only recently have there been attempts to develop methods for automatic grouped and hierarchical variable selection. Yuan and Lin (2006), Kim, Kim and Kim (2006) and Zhao, Rocha and Yu (2009) use generalizations of the LASSO (Tibshirani (1996)), that is they rely on the maximization of a penalized likelihood. The method of Yuan and Lin (2006) has been extended to logistic regression by Meier, van de Geer and Bühlmann (2008). These approaches are devised for grouped variable selection, and they accomodate hierarchical constraints by means of the definition of nested groups. The methods perform simultaneous shrinkage and selection, but can be hard to implement. Additionally, they rely on the maximization of a (penalized) likelihood over parameter spaces that may be non-convex. Another problem with LASSO-related methods is that they may not be consistent in model choice in certain situations (Meinshausen and Bühlmann (2006), Zou (2006)): the oracle penalty for optimal prediction is inconsistent for estimation of the true model unless an adaptive penalty is used, for example. Further, consistency is achieved only under conditions on the correlations between predictors inside and outside the true model. This kind of issue does not arise in the Bayesian framework, and the only condition we impose on the limiting correlation matrix is that it be positive definite.

The goal of this paper is to show that constrained model selection can be performed with a simple strategy in a Bayesian framework. Such a framework naturally embeds shrinkage of the estimates, and we prove it yields consistent model choice under weak conditions. By building on the stochastic search variable

selection (SSVS) approach of George and McCulloch (1993, 1997) we propose a Bayesian method for performing grouped, hierarchical, and anti-hierarchical variable selection. The aim of the stochastic search is to find "good models", rather than to determine the posterior distribution on the model space. The general agreement is that promising models may be sampled often (George and McCulloch (1993)) even when p is moderate. 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, Rocha and Yu (2009) provide an interesting Bayesian interpretation of their Composite Absolute Penalties.

The rest of the paper is as follows: in Section 2 we illustrate our strategy for constrained selection. Frequentist consistency in model choice, also in the general unconstrained framework, is shown in Section 3. We illustrate the method using simulations in Section 4 and, in the context of two data examples, in Section 5. Additional material, examples, and appendices can be found in the supplement at `http://www.stat.sinica.edu.tw/statistica`.

2. A Bayesian Model for Constrained Variable Selection

Bayesian model selection dates back at least to Atkinson (1978). There has been a huge amount of work on the subject since then, that we do not attempt to review; we just point the reader to Lahiri (2001), Chipman, George and McCulloch (2001), and the references therein. We focus in this paper on SSVS (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. We assume an observed response variable, Y, p predictors $X_1, \ldots, X_p$, some of which may be functions of the others, made on a sample of n subjects from a certain population. The key feature for performing SSVS is the introduction of a binary latent variable γ_j identifying whether the corresponding manifest variable should be included in the final model or not:

$$\beta_j \mid \gamma_j \sim (1-\gamma_j)N(0,\tau_{0j}^2) + \gamma_j N(0,\tau_{1j}^2); \tag{2.1}$$

with τ_{1j}^2 larger than τ_{0j}^2. Each model is then identified by a binary vector $\gamma = (\gamma_1, \ldots, \gamma_p)$, with prior probability $\pi(\gamma)$, in which the variables corresponding to non-zero components of γ are included and the other are excluded. The posterior probability of $\gamma_j = 1$ represents the posterior inclusion probability of variable X_j.

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

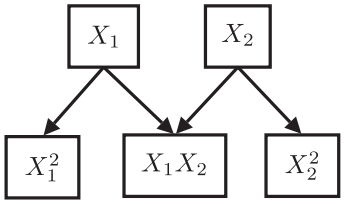

Figure 1. Illustration on a simple model

Example 1. Suppose we measure two covariates and a continuous response, and consider the possibility of including the square of each measurement and an interaction. The full model is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$. The set of constraints that preserves a hierarchical structure for a chosen submodel can be visualized in Figure 1, where a pointing arrow implies that corresponding variable cannot be included in the model without its parents.

More formally, in the constrained variable selection framework there is a proper subset $A \subseteq \{0,1\}^p$ such that, *a priori*, $P(\gamma \in A) = 1$. Since the applicability of SSVS does not depend on the functional form of the prior on γ, the solution reduces to building a distribution which reflects this prior knowledge about the structure of the model space. We propose to do this through the use of constraint indicator functions and a reparameterization.

We assume that the covariates are divided into g *disjoint* groups $G_1, \ldots, G_g$, and define indicators $\phi_k(j)=1$ if X_j is member of the kth group, $\phi_k(j)=0$ otherwise. In order to have a meaningful specification of disjoint groups, $\sum_k \phi_k(j)=1$ and $\sum_j \phi_k(j) \geq 1$. Variables that are not constrained to enter in groups go into a singleton. Hierarchical constraints are specified through indicators $\delta_j(i) = 1$ if the jth variable must be included in every model in which the ith is included, and zero otherwise. Define also indicators $\xi_j(i) = 1$, if the jth variable must be excluded from every model in which the ith is included, and zero otherwise. Note that ϕ, δ and ξ are fixed and pre-specified by the user.

After specification of ϕ, δ and ξ, we introduce new latent indicators η_k to identify whether the kth group is to be included in the final model or not. In our approach, *all* variables in a group are included simultaneously in the model when the corresponding indicator is equal to one. A convenient way of describing a prior on γ with the correct support is given by computing γ conditionally on η.

For instance, if there are only grouping constraints, the following equations must hold: $\gamma_j = \prod_{k=1}^{g} \eta_k^{\phi_k(j)}$. Since all but one $\phi_k(j) = 0$, $k = 1, \ldots, g$, γ_j will be set equal to the latent indicator η_k for which $\phi_k(j) = 1$.

If there is a hierarchical constraint between a father variable j_1 in group k_1 and variable j_2 in group k_2, the following equation must hold: $\gamma_{j_2} = \eta_{k_2}\eta_{k_1}$. In

fact, variable j_2 can be included only if *both* groups k_1 and k_2 are included in the model. A more general equation is given by $\gamma_{j_2} = \prod_{k=1}^{g} \eta_k^{\phi_k(j_2)} \prod_{j \neq j_2} \gamma_j^{\delta_j(j_2)}$. A similar approach can be taken in order to impose anti-hierarchical constraints. Hence, in the most general case, the following equation must hold:

$$\gamma_j = \Big(\prod_{j \neq i} (1-\gamma_i)^{\xi_i(j)} \gamma_i^{\delta_i(j)} \Big) \prod_{k=1}^{g} \eta_k^{\phi_k(j)}; \tag{2.2}$$

where for ease of notation we suppress the dependence of γ_j on η. A vector γ satisfying the constraints specified by the indicators ξ, δ and ϕ is obtained through equation (2.2), given η. For instance, $\gamma_j = 1$ only if $\prod_{k=1}^{g} \eta_k^{\phi_k(j)} = 1$, $\prod_{j \neq i} \gamma_i^{\delta_i(j)} = 1$ and $\prod_{j \neq i} (1-\gamma_i)^{\xi_i(j)} = 1$.

2.1. Constraint specification and solution of (2.2)

In this section we give some general guidelines on the specification of ξ, δ ,and ϕ, together with conditions for existence and uniqueness of a solution to (2.2). When it exists, a solution to (2.2) can be obtained through a simple iterative algorithm that is described at the end of the section.

When specifying the indicators the user should make sure that no hierarchial constraints could be replaced by grouping constraints: if we specify that X_i is a 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. This would lead to loss of uniqueness in the solution of (2.2). Further, the specification of the indicator functions $\delta_j(i)$ should be 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 redundant. Finally, the constraints should not be contradictory. 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.

In simple cases the parameterization (2.2) could be simply "exploded" and each element of the γ vector separately defined explicitly. 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 in the Supplementary material.

More formally, we put forward the following definitions.

Definition 1. A set of constraints given by a choice of ϕ, δ and ξ is called **minimal** if (i) there can be no switching between hierarchical and grouping constraints, and (ii) 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 order to formalize the idea of avoiding almost sure exclusion of certain groups, we introduce the concept of *compatibility*:

Definition 2. A set of constraints is called **compatible** if for each $j = 1, \ldots, p$ there exist a model in the subclass that includes X_j.

We expect the users to specify a minimal and compatible sets of constraints.

Examples 1(continued). In the model of Example 1, there are no grouping or anti-hierarchical constraints, while there are hierarchical constraints. This leads to $\phi_k(j) = 1_{k=j}$, 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. Each hierarchical constraint is put between variables belonging to different groups, and any removal would lead to the possibility of including at least one more model in the subclass of possible models. Hence the set is minimal. The full model is included in the collection, hence the set is compatible.

We can now show how to compute γ by solving (2.2) with an iterative method. Note that in general there can be more than one vector η leading to the same γ. We argue below that under minimality for given η there is only one vector γ which satisfies (2.2).

First, note that for each j, (2.2) is made of two factors that can be either zero or one. The second factor does not depend on the unknown γ, and can be computed directly. A starting solution can be given by setting $\gamma_j = \prod_{k=1}^{g} \eta_k^{\phi_k(j)}$. The elements which are set to zero coincide with those of the solution of (2.2) by construction. For the other elements, we can iterate the following step until no more changes are made to the vector γ:

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

If there are no anti-hierarchical constraints, the number of iterations is finite since changes to the starting solution involve only transitions from 1 to 0. If there are anti-hierarchical constraints it is straightforward to check that the number of iterations is finite as long as the set of constraints is compatible. If the set of constraints is compatible, anti-hierarchical and hierarchical constraints need not cycle, so that it cannot happen that setting some $\gamma_{j_1} = 0$ allows $\gamma_{j_2} = 1$, and the other way around. The resulting vector satisfies (2.2) by construction, which proves existence of the solution.

Uniqueness of the solution is guaranteed by minimality of the set of constraints. In fact, if there is no overlap between grouping and hierarchical constraints, the expression $\prod_{j \neq i} \gamma_i^{\delta_i(j)} = 1$, is *inactive* as long as the η_k corresponding to the *father* variables in the hierarchy are equal to one. As soon as a group in the hierarchy is excluded from the model, all the following are excluded by construction, so there can be only one solution. Anti-hierarchical constraints cannot lead to multiple solutions to (2.2) since a variable involved cannot be simultaneously in and out of a model for a given η.

There can be more than one solution to (2.2) if for instance some hierarchical constraints cycle (i.e., for some j_1, j_2 and j_3, $\delta_{j_1}(j_2) = \delta_{j_2}(j_3) = \delta_{j_3}(j_1) = 1$), leading to a spurious definition of a group. Such set of constraints would not be minimal by definition. If the set of constraints is not minimal and there is more than one solution to (2.2), the proposed strategy still works and leads to the solution corresponding to the model with the highest number of variables.

2.2. The model

We propose to fit the following hierarchical model:

$$\begin{cases} Y \mid \beta, \sigma^2 \sim N(\beta_0 + \sum \beta_k X_k, \sigma^2 I) \\ \sigma^2 \mid \eta \sim IG(\nu_\gamma/2, \nu_\gamma \lambda_\gamma/2) \\ \eta_k \sim \text{Bernoulli}(w_k) \\ \beta \mid \eta \sim N(0, \Gamma R \Gamma), \end{cases} \tag{2.3}$$

where $\Gamma = \text{diag}(\sqrt{\gamma_j \tau_{1j}^2 + (1-\gamma_j)\tau_{0j}^2})$, IG denotes the inverse gamma distribution, and R is a prior correlation matrix. The prior for β leads to a marginal prior as in (2.1) for each β_j. The only difference with the model proposed in George and McCulloch (1993) is that the latent variables γ_j enter into the model as the implicit solutions of (2.2). In the unconstrained case in which $G_j = \{j\}$, $j = 1, \ldots, p$ and $\delta_j = \xi_j = 0$ for any j, model (2.3) reduces to the model suggested in George and McCulloch (1993).

Our setting follows the approach of Bayesian variable selection in which the prior distribution of each β_j has a spike at zero. 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. The number of groups expected *a priori* to be included depends on the structure of the contraints; in the absence of hierarchical and anti-hierarchical contraints it is easily seen to be equal to $\sum_k w_k$. Larger models are easily penalized by small values of w_k. The parameter γ_j is equal almost surely (conditionally on η) to a function of η. Marginally, its prior is 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 one to give zero prior probability to models that do not satisfy

the constrains in a simple and natural way. Choice of parameters for the priors is discussed in the supplementary material.

2.3. MCMC Sampling

SSVS can be performed with the use of classical MCMC methods (Robert and Casella (1999)), namely by the implementation of a simple Gibbs sampler. In our experience, in the constrained framework no additional issues with respect to the unconstrained SSVS seem to emerge. In many cases the Gibbs sampling scheme described in this section will be enough to obtain reliable posterior summaries for model selection. When this is not the case, sampling strategies as described in Madigan and York (1995), Geweke (1996), George and McCulloch (1997) and Hans, Dobra and West (2007) may be used. A particularly 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 multimodality of the posterior (as in our case for the marginal posteriors of the β parameters). In Section 3.2 of the supplementary material we describe a sampling strategy especially devised for large model spaces.

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

$$\beta \mid Y, X, \sigma^2, \eta \sim N((X'X + D^{-1}R^{-1}D^{-1})^{-1}X'Y, \sigma^2(X'X + D^{-1}R^{-1}D^{-1})^{-1}), \tag{2.4}$$

where $D = \text{diag}(\sqrt{\gamma_j\tau_{1j}^2 + (1-\gamma_j)\tau_{0j}^2}/\sigma)$. When considering a large number of transformations and interactions, the number of prospective predictors p can get much larger than n. This is not a problem since $(X'X + D^{-1}R^{-1}D^{-1})$ is positive definite, hence invertible, for any p. The full conditional for the variance is:

$$\sigma^2 \mid Y, X, \beta, \eta \sim IG(\frac{n+\nu_\gamma}{2}, \frac{\nu_\gamma\lambda_\gamma + |Y - X\beta|^2}{2}), \tag{2.5}$$

while for the latent variables η we have

$$\eta_k \mid \beta, \sigma^2 \sim \text{Bernoulli}(\frac{w_k a}{w_k a + (1-w_k)b}), \tag{2.6}$$

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 in which the kth component has been removed. It is interesting to note that the full conditional of η_k does not depend on Y, since Y depends on η only through the vector β (George and McCulloch (1993)). If we do not let ν_γ and λ_γ depend on γ, the parameter of the Bernoulli in (2.6) further simplyfies to $f(\beta \mid \eta_{-k}, \eta_k = 1)w_k/(f(\beta \mid \eta_{-k}, \eta_k = 1)w_k + f(\beta \mid \eta_{-k}, \eta_k = 0)(1-w_k))$. The components of the vector η are sampled

sequentially; in our experience, randomly changing the order at each sweep of the chain mildly improves the mixing properties of the Gibbs sampler. Finally, γ_j is computed from η according to (2.2).

2.4. Alternative approaches to model choice

In order to perform model choice we exploit the posterior for the γ parameters, $\pi(\gamma \mid Y)$. In this section we assume the posterior has already been approximated through MCMC sampling. In this case, $\pi(\gamma \mid Y)$ is simply estimated as the relative frequency each model is sampled, and equivalently $\pi(\gamma_j \mid Y)$ is the relative frequency each variable is sampled. Inference on η_k can be performed similarly, but it is not usually of interest.

Common approaches to model choice select the one receiving highest posterior probability. 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 0.5 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 from a predictive point of view 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 the collection. In that case the model receiving highest posterior probability is to be selected.

Use of the median model is advantageous also from a computational point of view since it only involves evaluation of the marginal probabilities $\pi(\gamma_j \mid Y)$, and not of the joint $\pi(\gamma \mid Y)$. Evaluation of the latter in principle requires enumerating all sampled configurations of the entire vector, which may be large even when p is moderate.

Even if we focus here on model choice, we also give some consideration to 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 still be useful since a model that is known *a priori* not to hold should receive zero posterior probability.

2.5. Extension to generalized linear models

We now discuss an extension to Generalized Linear Models (GLM), see McCullagh and Nelder (1989). In a GLM the response is $Y \sim \exp\{y\theta - b(\theta)/a(\phi) + c(y,\phi)\}$, with parameters θ and ϕ, known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$, and an assumed linear relation $g(E[Y \mid X]) = \beta_0 + \sum \beta_k X_k$, 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 as needed. As before, a Gibbs sampler can be set up to simulate from the posteriors. As pointed out by Dellaportas and Smith (1993), the adaptive rejection method can be used to approximate the posterior when $g(\cdot)$ is a canonical link function (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, the adaptive rejection Metropolis sampling of Gilks, Best and Tan (1995) can be used.

3. Frequentist Properties

In this section we prove certain consistency results for Bayesian variable selection. We point out that these results hold for SSVS both in the constrained and unconstrained cases. We use the shorthand notation of M_0 for the γ vector corresponding to the true model and write $M_{me} \mid Y$ for the vector corresponding to the posterior median model.

Theorem 1. *Assume $(X'X)/n \to C$, where C is positive definite, and assume the true and median models are included in the collection of possible models. Fix $w_k > 0$, $\tau_{0j}^2 < \tau_{1j}^2$, and*

$$(1 - w_k)\tau_{1j}^2 > w_k \tau_{0j}^2 \tag{3.1}$$

for all $k = 1, \ldots, g$ and $j = 1, \ldots, p$. Let ν_γ and λ_γ 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 ijth element of R^{-1} and β^* is the vector of true parameters. Then $\lim_{n\to\infty} \Pr(M_{me} = M_0 \mid Y) = 1$, and if $\max_j \tau_{0j} \xrightarrow{n} 0$, $\lim_{n\to\infty} \Pr(M_0 \mid Y) = 1$.

Proof. Proof in Appendix A in the supplementary material.

Theorem 1 has it 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 if τ_0 is infinitesimal (or exacly equal to zero). An equivalent expression of (3.1) is $0 < w_k < \min_j \tau_{1j}^2/(\tau_{1j}^2 + \tau_{0j}^2)$, 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 (3.1) 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 that could be removed with some restrictions on the magnitude of the prior correlations. Nevertheless, it is straightforward to check that if R is the identity matrix the condition is satisfied for any finite n, while $R \propto (X'X)^{-1}$ asymptotically suffices for Theorem 1. For consistency of the model with highest posterior probability, we need to let τ_{0j} decrease to zero. As pointed out above, there are no problems in setting $\tau_{0j} = 0$ for all n if the appropriate sampling algorithm is used.

It is particularly surprising that the results hold without further conditions on X. For prediction for instance, orthogonality or other restrictions are needed to prove that the median model is optimal. The theorem provides weaker results in many senses. Consistency of the model receiving highest posterior probability has been long known 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 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, however, consistency results for the median model are new also for the case of unconstrained variable selection.

4. Simulations

We did a small simulation study in order to check on the ability of the constrained setting to pick the right model, and on prior sensitivity. First, we generated six covariates $X_1, \ldots, X_6$ from standard normals, with response from

$$Y = 1.5X_1 + 2X_2 + X_3 - 1.5X_2X_3 + \varepsilon, \tag{4.1}$$

where $\varepsilon \sim N(0, 9)$. The sample size was $n = 250$. We allowed inclusion of any of the 6 available covariates, and any of the 15 possible bivariate interactions. We imposed constraints so as to sample only models respecting a hierarchical structure. We set $w_k = 0.5$, R to be an identity matrix, $\tau_0^2 = 0.0625$, and $\tau_1^2 = 1,000$. We generated the data, used a burn-in of 5,000 sweeps, and let the sampler run for another 3,000 iterations which we used for model choice. We selected the median model. We repeated the operation for $B = 300$ iterations, and report the proportion of times the strategy led to the correct model choice. Proportions of correct model choice are reported in the Scenario 1 row in Table 1, together with the Monte Carlo standard error. In Scenario 2 we still simulated from model (4.1), but also imposed an anti-hierarchical constraint between X_4 and X_6. In Scenario 3 we imposed two anti-hierarchical constraints, one between X_5 and X_2, and another between X_4 and the interaction X_2X_3; note that this implies also an anti-hierarchical constraint 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 raised the rate of correct decision, while anti-hierarchical constraints between variables not included in the true model did not seem to have a significant effect. In Scenario 4 we used the same constraints as in Scenario 1, but took $X_4 = X_5 + \varepsilon_2$, where $\varepsilon_2 \sim N(0, 0.01)$. This introduced a strong collinearity in the design matrix, but did not seem to alter the ability of the algorithm to choose the right model. In Scenario 5 we used the same constraints as Scenario 1, and also imposed grouping constraints between X_2 and X_3, and between X_5 and X_6. In Scenario 6 we simulated from model (4.1), but considered 25 possible covariates and all their bivariate interactions, ending up with $p = 325 > n = 250$. Since the model space was now larger, while we still used a burn-in of 5,000 iterations, we let the sampler run for 15,000 more iterations. Having a much larger model space decreased the probability of correct model selection. Nevertheless, with $\tau_1 = 100$ and $\tau_0 = 0.05$ (Scenario 7), the probability of correct model selection increased (See also Table 2 below).

If the constraints are misspecified and the correct model is almost surely excluded from the collection of possible models, constrained SSVS fails to select the true model. On the other hand, it is well-known that a model "close" to the correct one is chosen (Berk (1966), Dmochowski (1996)). To illustrate this, we simulated as in Scenario 1 with an anti-hierarchical constraint between X_2 and X_1. The correct model was now obviously never sampled. On the other hand, the median model was $\beta_0 + \beta_2 X_2 + \beta_3 X_3 - \beta_{23} X_2 X_3$ in 280 out of 300 simulated data sets.

For a comparison we used unconstrained SSVS on the same data, with the same priors and MCMC tuning. The proportion of correct model selection is seen to decrease considerably, as reported in the right panel of Table 1. In particular when $p > n$, classical SSVS is led to almost always include too many variables. In summary, putting (the right) constraints narrows the search for the true model and increases the probability of correct model selection.

Finally, in order to evaluate the effect of the choice of prior parameters, we simulated from Scenario 1 but tried different values for τ_1, τ_0 and w_k. For each combination of τ_1 and τ_0 with $w_k = 0.5$ we generated $B = 100$ data sets, and we report in Table 2 the proportion of times constrained SSVS with use of the median model was led to choose the correct model. The last line of Table 2 shows the results for τ_0 and τ_1 as in Scenario 1, but with different values for w_k (remember that setting $w_k > 0.5$ violates the conditions of Theorem 1). A certain degree of dependence on prior inputs is well known in Bayesian variable selection, and confirmed by the simulation; but a reasonable range of choices led to choose the correct model with high probability. A general guideline is not possible since

Table 1. Proportion of correct model selection under different scenarios for SSVS

	Constrained SSVS		Unconstrained SSVS	
Scenario	Correct	Monte Carlo Standard Error	Correct	Monte Carlo Standard Error
Scenario 1	0.93	0.014	0.57	0.028
Scenario 2	0.92	0.015	0.57	0.028
Scenario 3	0.96	0.011	0.57	0.028
Scenario 4	0.93	0.015	0.56	0.028
Scenario 5	0.94	0.013	0.57	0.028
Scenario 6	0.66	0.027	0.03	0.010
Scenario 7	0.91	0.016	0.05	0.012

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

	$\tau_1^2 = 5$	$\tau_1^2 = 10$	$\tau_1^2 = 100$	$\tau_1^2 = 200$	$\tau_1^2 = 500$	$\tau_1^2 = 1,000$
$\tau_0^2 = 0.1$	1.00	1.00	0.90	0.83	0.75	0.59
$\tau_0^2 = 0.05$	0.99	1.00	1.00	1.00	1.00	0.98
$\tau_0^2 = 0.02$	0.98	0.98	1.00	1.00	1.00	0.99
$\tau_0^2 = 0.005$	0.97	0.99	0.99	0.99	1.00	1.00
$\tau_0^2 = 0.001$	0.96	0.97	0.98	0.98	0.99	1.00
w_k	0.1	0.25	0.4	0.6	0.75	0.85
	0.78	0.85	0.96	0.94	0.93	0.90

this "reasonable range" depends heavily on the data at hand. It has been noted in the literature that it is better to have a dense model space, with constituent models close together (Gustafson and Lefebvre (2008)). If the prior variances are too separated, the different models are too far apart and a correct mixing of the chain is difficult, at least with simple sampling schemes. We suggest tuning prior inputs by exploring the parameters for the full conditionals of η_k as given in (2.6), and making sure they move slowly as a function of β. With bad values it may happen that parameters in (2.6) jump between the extremes of the interval $[0, 1]$. Further discussion of the choice of prior parameters can be found in the supplementary material.

5. Data Examples

In this section we provide an example with a continuous response and another with a large and complex model space and a binary response. An extended version of the latter, together with a brief discussion of two additional examples, can be found in the supplementary material.

5.1. Birthweight data

Consider the birthweight data set from Hosmer and Lemeshow (1989). We have $n = 189$ observations collected by the Baystate Medical Center, Springfield,

Massachusetts, during 1986. Response is weight at birth, and there is information on mother's age, weight at last menstrual period, race (white, black, other), smoking status during pregnancy, and number of previous premature labours, hypertension in the past, uterine irritability, number of physician visits during the first trimester. We considered 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 adopted a corner point parameterization for race.

To fix the ideas, we describe how the constraints were specified for the variables `race` and `weight`. For `race`, we used `white` as baseline. First we specified grouping constraints by setting $\phi_1(\texttt{black}) = \phi_1(\texttt{other}) = 1$. All the other variables belong to a different group, so for instance we set $\phi_2(\texttt{weight}) = \phi_3(\texttt{smoke}) = \phi_4(\texttt{weight}^2) = 1$, and so on. We then specified separate hierarchical constraints by setting $\delta_{\texttt{weight}}(\texttt{weight}^2) = \delta_{\texttt{weight}^2}(\texttt{weight}^3) = 1$, $\delta_{\texttt{black}}(\texttt{weight}*\texttt{black}) = \delta_{\texttt{weight}}(\texttt{weight}*\texttt{black}) = 1$, and so on. After a burn-in of 50,000 iterations we let the Gibbs sampler run for another 50,000.

One question of interest is whether the constrained framework significantly modifies the correlation structure among the parameters with respect to the unconstrained framework. In particular, if parameters are highly correlated the Gibbs sampler may not be the most efficient choice for sampling from the posterior. With these data we provide mild evidence that this is not the case. After sampling, we computed the correlation matrix for the regression parameters. We repeated the operation after sampling in an unconstrained SSVS framework. In the first case the largest eigenvalue of the correlation matrix was 3.671, and the smallest 0.0002. In the second case the largest eigenvalue was 3.634, and the smallest 0.0003. Since there usually are many more samples from the posterior than parameters, this simple check can always be done. Should the parameters in the constrained framework be much more correlated than the parameters in the unconstrained framework, we suggest better tuning of the prior correlations, of τ_0 and τ_1, or the use of another sampling strategy.

The median model and the model with highest posterior probability coincide:

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

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 was high (872). In this example Bayesian model averaging may be a better choice if the goal is prediction. In

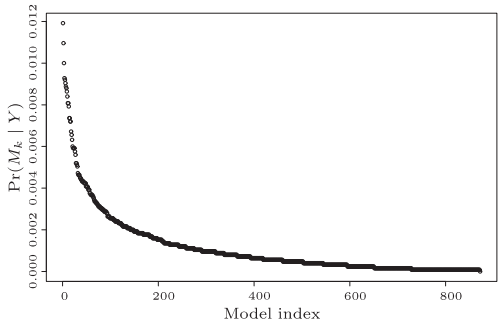

Figure 2. Posterior probability of sampled models (decreasing order) for the Birthweight data set.

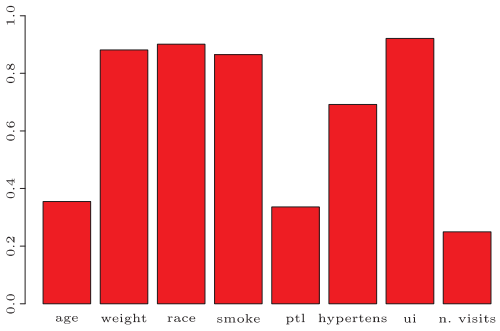

Figure 3. Posterior probability of original variables for the Birthweight data set. `ui`: uterine irritability, `ptl`: number of premature labours.

Figure 3 we show the posterior inclusion probability of each variable, excluding transformations and interactions for reasons of space. This latter plot gives more convincing evidence in favour of the chosen model.

The same data were analyzed by Yuan and Lin (2006). The selected model substantially coincide with the one suggested by them, and we also agree in identifying the number of visits as the least important covariate (posterior inclusion probability: 0.25), and uterine irritability as the most important (posterior inclusion 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 very likely be excluded, having marginal posterior inclusion probability of 0.003. Finally, we speculate that the Bayesian hierarchical model selects hypertension because of the presence of an interaction with race. 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.

Table 3. Average MSE on the test set with its Monte Carlo standard error and average number of variables selected by different methods for the Birthweight data. The results are based on 1,000 random partitions of the data in a training set of 151 observations and a test set of 38 observations.

Method	MSE	Monte Carlo Standard Error	# Variables
constrained SSVS	444116.6	3019.99	8.08
grouped LASSO	459114.8	3309.98	10.72
unconstrained SSVS	459719.0	2898.14	6.77
forward stepwise	468421.5	3300.72	9.36

In order to evaluate the predictive performance of our method, we split the data set into a training sample of $n = 151$ observations and a test sample of the remaining 38. We used the same constraints as in Yuan and Lin (2006), with grouping constraints among the polynomial transformations up to the third order for the numerical variables, and among the coefficients for the corner point reparameterization of race. We used our method and competitors for model selection and we predicted the responses on the test set with the chosen model. We repeated the operation $B = 1,000$ times and report in Table 3 the average MSE, together with Monte Carlo standard error, and the average number of selected variables. Constrained SSVS stands for our method, unconstrained SSVS for the classical stochastic search variable selection. In both cases the final choice was based on the median model. We compared the Bayesian methods with two frequentist approaches, the classical forward stepwise selection and the grouped LASSO. For the grouped LASSO of Yuan and Lin (2006) we used a grid of 50 values for the penalization parameter, and chose the result optimizing their approximate C_p criterion. Despite not being optimized for prediction SSVS, both in the constrained and unconstrained version, seemed competitive with respect to the frequentist criteria.

5.2. Spam data

In order to illustrate the potential of our method with many variables and many constraints, we show the 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 from 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 in the supplementary material.

It is natural to expect high-order interactions between the occurrence of certain words in this data set. We proceeded by randomly splitting the data set into

a training set of 3,221 observations and a test set of the remaining 1,380. We considered 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 was 5,940. As usual we did not include a power of any order without all the preceeding, nor an interaction without the original variables. There was a very large number of complex constraints, which could not be easily exploded. Nevertheless, the hierarchical constraints were easily specified by forming a 5,940 by 5,940 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 untransformed); and so on. In order to penalize more complex models, we set $w_k = 0.05$. The prior variances were $\tau_1 = 5$ and $\tau_0 = 0.3$. Due to the large number of variables, the Gibbs sampler could not be expected to converge without a very large number of iterations, so we used instead a special sampling scheme, whose description can be found in the supplementary material.

The resulting 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. There is very little uncertainty about the chosen model (see supplementary material for a deeper discussion about this point), which is also interpretable. For instance, the semicolon was discarded, while name of the owner of the mailbox was 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 were "hp" and "meeting". With very high probability we also included the words "order", "technology", and "000", that are common words in spam emails. Many words interact with the name of the owner and the word "hp". Other interactions formed part of a sentence, like: "our" and "meeting". When the two words are used together, it is less likely the mail is spam and the negative coefficient for the interaction catches this feature. Another interesting interaction is between "hp" and "technology". 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 to be spam, reflected on the negative coefficient of the interaction. Not surprisingly, even if "000" is very important in the model, there are only six interactions with this variable.

Finally, we used the selected model for prediction on the test set. The results are shown in Table 4, 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

Table 4. 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%

that no variable selection is available for the k-nn methods. For all the other methods, a (constrained or unconstrained) SSVS is used. The predictive 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 increased from 90.4% to 91.6%). If the transformations are used without constraints, the prediction performance was not as good, likely due to over-adaptation to the training set. Moreover, the resulting model was not easily interpretable and not as parsimonious, since it used 711 variables. The logistic model with hierarchical contraints can be used not only for prediction, but also for explaining why an email is spam.

Acknowledgements

The author is grateful to an associate editor and two anonymous referees for kind suggestions that led to improvements of the paper. Acknowledgements go also to Prof. Pierluigi Conti for careful checking of the proofs, and to Prof. Luca Tardella for advice on an earlier draft.

References

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

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

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection, *Ann. Statist.* **32**, 870-897.

Berk, R. (1966). Limiting behavior of posterior distributions when the model is incorrect, *Ann. Math. Statist.* **37**, 51-58.

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

Cholongitas, E., Marelli, L., Shusang, V., Senzolo, M., Rolles, K., Patch, D. and Burroughs, A. (2006). A systematic review of the performance of the model for end-stage liver disease (MELD) in the setting of liver transplantation, *Liver Transplantation* **12**, 1049-1061.

Clyde, M. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics* **6** (Edited by J. Bernardo, J. Berger, A. Dawid and A. Smith), Oxford Press, 157-185.

Cordell, H. (2002), Epistasis: what it means, what it doesn't mean, and statistical models to detect it in humans, *Human Molecular Genetics* **11**, 2463-2468.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *IEEE Trans. Inform. Theory* **IT-13**, 21-27.

Dellaportas, P. and Smith, A. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling, *Appl. Statist.* **42**, 443-459.

Dmochowski, J. (1996), Intrinsic priors via Kullback-Leibler geometry. In *Bayesian Statistics* **5** (Edited by J. Bernardo, J. Berger, A. Dawid and A. Smith), Oxford Press, 543-549.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.* **88**, 881-889.

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

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

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

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337-348.

Gustafson, P. and Lefebvre, G. (2008). Bayesian multinomial regression with class-specific predictor selection. *Ann. Appl. Statist.* **2**, 1478-1502.

Hans, C., Dobra, A. and West, M. (2007). Shotgun search for "large p" regression, *J. Amer. Statist. Assoc.* **102**, 507-516.

Hoeting, J., Madigan, D., Raftery, A. and Volisnky, C. (1999). Bayesian model averaging: a tutorial. *Statist. Sci.* **14**, 382-417.

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

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

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

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

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

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data, *Internat. Statist. Rev.* **63**, 215-232.

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

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70**, 53-71.

Meinshausen, N. and Bühlmann, P. (2006). Variable selection and high-dimensional graphs with the Lasso. *Ann. Statist.* **34**, 1436-1462.

Meyer, M. and Laud, P. (2002). Predictive variable selection in generalized linear models. *J. Amer. Statist. Assoc.* **97**, 859-871.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical Methods*. Springer-Verlag Inc.

Scheffè, H. (1963), *The Analysis of Variance*. Wiley, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37**, 3468-3497.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

"Sapienza" University of Rome, Piazzale Aldo Moro 5, 00185, Rome, Italy.

E-mail: alessio.farcomeni@uniroma1.it

(Received October 2007; accepted April 2009)