UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

Working Paper M09/04

Methodology

Two Level Stochastic Search Variable Selection In GLMs With Missing Predictors

Robin Mitra, David D. Dunson

Abstract

Stochastic search variable selection (SSVS) algorithms provide an appealing and widely used approach for searching for good subsets of predictors, while simultaneously estimating posterior model probabilities and model-averaged predictive distributions. This article proposes a two-level generalization of SSVS to account for missing predictors, while accommodating uncertainty in the relationships between these predictors. Bayesian approaches for allowing predictors that are missing at random require a model on the joint distribution of the predictors. We show that predictive performance can be improved by allowing uncertainty in the specification of this model. The methods are illustrated through simulation studies and analysis of an epidemiologic data set.

Two Level Stochastic Search Variable Selection in GLMs with Missing Predictors

Robin Mitra

Southampton Statistical Sciences Research Institute,

University of Southampton,

Southampton, SO17 1BJ, UK

R.Mitra@soton.ac.uk

David B. Dunson

Department of Statistical Science, Duke University,

Box 90251, Durham, North Carolina 27708, U.S.A.

dunson@stat.duke.edu

Abstract

Stochastic search variable selection (SSVS) algorithms provide an appealing and widely used approach for searching for good subsets of predictors, while simultaneously estimating posterior model probabilities and model-averaged predictive distributions. This article proposes a two-level generalization of SSVS to account for missing predictors, while accommodating uncertainty in the relationships between these predictors. Bayesian approaches for allowing predictors that are missing at random require a model on the joint distribution of the predictors. We show that predictive performance can be improved by allowing uncertainty in the specification of this model. The methods are illustrated through simulation studies and analysis of an epidemiologic data set.

Key words: Missing at random; Model averaging; Multiple imputation; Stochastic search; Subset selection; Variable selection;

1 Introduction

In regression, one issue that is routinely encountered is how to select a subset of the available predictors that are important in explaining the response. In many fields, this variable selection problem is faced in essentially every study that is conducted. For example, in epidemiologic studies of exposure-disease relationships, investigators typically collect information on multiple potential risk factors and confounding variables. Clearly, problems can be encountered if all these variables are included as predictors, so epidemiologists tend to discard covariates that do not have a significant impact on disease risk, unless these covariates are the primary exposures of interest or there is strong prior knowledge that they should be included.

Stepwise selection is the most widely-used automated algorithm for selecting variables to include in a regression model, with many variants possible depending on the starting model, the manner in which variables are added or deleted, and the criteria for deciding whether a predictor significantly improves goodness-of-fit. For example, forward selection sequentially adds predictors, keeping those that improve the AIC, BIC, or have p-values in a likelihood ratio, Wald or score test below some pre-specified threshold. For generalized linear models (GLMs), the order in which variables are added and the criteria used can have a substantial impact on the final subset of variables that are selected. In addition, basing inferences on the model selected from a stepwise procedure without accounting for uncertainty in the selection process can lead to highly misleading results. For example, there will be a greatly inflated type I error rate and the parameter estimates will be biased away from zero, particularly if there are many candidate predictors.

A number of strategies have been proposed to address such problems, with the focus in this article on Bayesian model averaging approaches allowing for missing predictors. In the Bayesian paradigm, one can assign posterior probabilities to each of the models in a list of *a priori* plausible models. To avoid uncertainty in model selection, one can then average over models in the list using posterior probability weights in performing inferences and predictions. In terms of prediction, Bayesian model averaging has been shown to have better performance compared with using any single model (Raftery *et al.*, 1997). For a recent review of Bayesian model averaging, refer to Clyde and George (2004).

In variable selection problems, the list of models under consideration corresponds to the 2^p possible subsets of a set of p candidate predictors. Clearly, the number of models rapidly becomes enormous as p increases, so there is a need for efficient methods for searching for high posterior probability models, while also estimating posterior model probabilities and the posterior distributions for the coefficients in each model. A widely used strategy for addressing this problem is to embed all the models in a full model containing all the predictors, and then allow predictors to drop out by choosing a mixture prior for the coefficients with one component concentrated at zero (Mitchell and Beauchamp, 1988). One can then use a Gibbs sampling algorithm for simultaneous model search and posterior computation, with such an approach referred to as stochastic search variable selection (SSVS) (George and McCulloch, 1993, 1997).

When missing values are present in the covariates SSVS algorithms cannot be applied directly. One commonly used strategy is to discard subjects with any missing predictors (complete case analysis), but this can be a sizeable proportion of the subjects in variable selection contexts, as one would need to discard subjects with missing values in any of the candidate predictors. Further, when missing patterns are not MCAR this approach can lead to biased inferences. Bayesian models can easily accommodate missing predictors by placing a joint model on the distribution of the predictors and then imputing the missing values within an MCMC algorithm. In the variable selection setting, with the response and predictors following a multivariate normal distribution, such an approach was implemented by Yang *et al.* (2006). This article addresses a much broader class of models involving mixed categorical and continuous variables, while also allowing model selection for the predictor component.

Outside of the variable and model selection context, a standard approach for specifying the joint distribution of the predictors, while allowing these predictors to have different measurement scales, is to choose a sequence of GLMs (see, for example, Lipsitz and Ibrahim (1996); Ibrahim *et al.* (1999)). However, following such an approach one faces uncertainty in how to specify the GLMs for X_1, X_2 given X_1, X_3 given X_1, X_2, etc. This is essentially another level of variable selection, so it is natural to allow uncertainty in this component of the model as well. This article proposes a two-level SSVS approach to allow uncertainty in the exact form of the imputation models for the missing covariate data. By allowing more parsimonious modeling of the joint predictor distribution through model averaging, we anticipate an improvement in predictive performance.

Section 2 briefly reviews the Bayes approach to model uncertainty in variable selection, and describes how to accommodate missing predictors in this paradigm. Section 3 presents the priors and models used to implement a two-level SSVS algorithm. Section 4 provides theoretical support for the approach. Section 5 illustrates performance of the method through simulation studies. Section 6 presents an application to an epidemiologic study, and Section 7 concludes with a discussion.

2 Two Level Variable Selection

2.1 Review of Bayesian Variable Selection

Suppose data for subject i $(i = 1, \ldots, n)$ consist of a response y_i and a vector of candidate predictors, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ denote a vector of predictor inclusion indicators, with $\gamma_j = 1$ denoting that the j^{th} element of $\boldsymbol{x_i}$ should be included in the regression model for the response and $\gamma_j = 0$ otherwise. Then, we focus on the case in which the conditional likelihood of y_i given $(\boldsymbol{x}_i, \boldsymbol{\gamma})$ belongs to an exponential family with scale parameter τ and location parameter $\mu_i = E(y_i|\boldsymbol{x}_i, \boldsymbol{\gamma})$, with $g(\mu_i) = \boldsymbol{x}'_{\boldsymbol{\gamma}i}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, where $\boldsymbol{x}'_{\boldsymbol{\gamma}i} = \{x_{ij}, j : \gamma_j = 1\}$ is the subset of predictors included in the model indexed by $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\beta_{\boldsymbol{\gamma}1}, \ldots, \beta_{\boldsymbol{\gamma}p_{\boldsymbol{\gamma}}})'$ denotes the coefficients for model $\boldsymbol{\gamma}$, $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ is the number of predictors in model $\boldsymbol{\gamma}$, and $g(.)$ is a known link function.

Hence we have defined a typical variable selection problem in the setting of a generalized linear model (GLM). There are 2^p possible indicator vectors $\boldsymbol{\gamma}$, with the model space corresponding to these different possibilities denoted by $\boldsymbol{\Gamma}$. A Bayesian formalization of the variable selection problem requires a prior for $\boldsymbol{\gamma}$ with support on $\boldsymbol{\Gamma}$, as well as a prior on the coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ for each $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$. The posterior probability allocated to model $\boldsymbol{\gamma}$ is then defined via Bayes rule as:

$$p(\gamma|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{\gamma})p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}^* \in \boldsymbol{\Gamma}} p(\boldsymbol{\gamma}^*)p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\gamma}^*)}, \tag{1}$$

where $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\gamma}) = \int \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_{\boldsymbol{\gamma}i}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau) dp(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau)$ is the marginal likelihood of the data under model $\boldsymbol{\gamma}$, $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau)$ is the prior on the coefficients and scale parameter in model $\boldsymbol{\gamma}$, and $p(\boldsymbol{\gamma})$ is the prior probability of model $\boldsymbol{\gamma}$.

For linear regression models and conjugate priors, the marginal likelihood under each model is available in closed form and the main practical issues that arise are (1) how to choose $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau)$, noting that model selection is sensitive to this choice (Liang *et al.*, 2008); and (2) how to efficiently search the model space given that the number of subsets increases rapidly with p. For non-normal GLMs, the Laplace approximation to the marginal likelihood can be used (Raftery, 1996).

2.2 Bayes Variable Selection with Missing Predictors

Now consider the common setting in which only a subset of predictors are observed. In particular let $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{ip})'$ denote a vector of missingness indicators specific to subject i with $m_{ij} = 1$ denoting that the j^{th} predictor is missing. In this setting, the approach described in section 2.1 cannot be applied directly.

Using the formulation described in Little and Rubin (2002), we consider the full marginal likelihood under the model $\boldsymbol{\gamma}$: $p(\boldsymbol{y}, \boldsymbol{M}|\gamma, \boldsymbol{\phi}, \boldsymbol{X_{obs}})$, where $\boldsymbol{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)'$ is the $n \times p$ matrix of missingness indicators for all subjects, $\boldsymbol{\phi}$ are parameters characterizing the likelihood of the missingness indicators, $\boldsymbol{X}_{obs} = \{x_{ij}, i = 1, \ldots, n, j : m_{ij} = 0\}$ are the observed predictor values, and $\boldsymbol{X}_{mis} = \{x_{ij}, i = 1, \ldots, n, j : m_{ij} = 1\}$ are the missing predictor values. We express this joint likelihood for $\boldsymbol{y}$ and $\boldsymbol{M}$ given $\boldsymbol{\phi}$ and the observed predictors in a selection model form as follows:

$$p(\boldsymbol{y}, \boldsymbol{M}|\gamma, \boldsymbol{\phi}, \boldsymbol{X}_{obs}) = \int p(\boldsymbol{y}|\boldsymbol{X}, \gamma)p(\boldsymbol{M}|\boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{X})p(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \gamma)d\boldsymbol{X}_{mis}.$$

When predictors are MAR, $p(\boldsymbol{M}|\boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{X}) = p(\boldsymbol{M}|\boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{X}_{obs})$. In addition, when the parameters governing the observed data likelihood and the missing data mechanism are distinct, in that the prior distributions on these parameters are independent, the missing data mechanism is ignorable and we can base inferences on the observed data likelihood,

$$p(\boldsymbol{y}|\boldsymbol{X}_{obs}, \boldsymbol{\gamma}) = \int p(\boldsymbol{y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau)p(\boldsymbol{X}_{\boldsymbol{\gamma}mis}|\boldsymbol{X}_{\boldsymbol{\gamma}obs})d\boldsymbol{X}_{\boldsymbol{\gamma}mis}dp(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau), \tag{2}$$

where $\boldsymbol{X}_{\boldsymbol{\gamma}obs} = \{x_{ij}, i = 1, \ldots, n, j : (1 - m_{ij})\gamma_j = 1\}$ and $\boldsymbol{X}_{\boldsymbol{\gamma}mis} = \{x_{ij}, i = 1, \ldots, n, j : m_{ij}\gamma_j = 1\}$.

We treat the missing covariate data $\boldsymbol{X}_{\boldsymbol{\gamma}mis}$ in model $\boldsymbol{\gamma}$ as nuisance parameters to be integrated out of the likelihood. In this way we can estimate the posterior probability of model $\boldsymbol{\gamma}$ using equation (1), but with the marginal likelihood defined conditionally on the observed data. In order for (2) to be well defined, we require a probability model for the joint distribution of the predictors, so that one can obtain the conditional likelihood of $\boldsymbol{X}_{mis}$ given $\boldsymbol{X}_{obs}$. We initially describe such a model without allowing for uncertainty in the choice.

In particular, following common practice in the literature on missing predictors having mixed measurement scales, we use the factorization:

$$p(\boldsymbol{X}) \;=\; p(\boldsymbol{x}_1)\prod_{j=2}^{p} p(\boldsymbol{x}_j|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{j-1}), \tag{3}$$

where $p(\boldsymbol{x}_j|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{j-1}) = \int \prod_{i=1}^{n} p(x_{ij}|x_{i1},\ldots,x_{i,j-1},\boldsymbol{\theta}_j,\kappa_j)dp(\boldsymbol{\theta}_j,\kappa_j)$ is characterized as a distribution in the exponential family with dependence on previous predictors modeled via a GLM with $\boldsymbol{\theta}_j,\kappa_j$ the regression coefficients and dispersion parameter, respectively, in the jth GLM in the sequence, with $p(\boldsymbol{\theta}_j,\kappa_j)$ the prior distribution. We also model $\boldsymbol{x}_1$ to be in the exponential family conditional on location and scale parameters θ_1 and κ_1. Then denote $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j, j = 1,\ldots,p\}$ and $\boldsymbol{\kappa} = (\kappa_1,\ldots,\kappa_p)'$ to be the set of regression coefficients and dispersion parameters characterizing the joint distribution of the predictors.

One could run an MCMC algorithm to generate samples from the conditional distribution of $\boldsymbol{X}_{mis}$ given $\boldsymbol{X}_{obs}$, $\boldsymbol{y}$ and $\boldsymbol{\gamma}$. These samples could be used to fill in the missing predictors at each sampling step of an SSVS analysis that accounts for uncertainty in the predictors to be included in the response model. However, this approach would not allow uncertainty in specification of the models characterizing (3).

2.3 Variable Selection for the Missing Data Model

When the number of predictors is large, questions arise in specification of each of the regression models, $p(\boldsymbol{x}_j|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{j-1},\boldsymbol{\theta}_j,\kappa_j)$. We are faced with essentially the same issues that motivate variable selection in our 'top level' model relating the response to the predictors;

in particular, there could be sparse relationships between variables and so it may not be necessary to include all $j-1$ predictors in our model for $\boldsymbol{x}_j$.

A natural extension is to perform variable selection within each of the conditional regression models characterizing the joint distribution of the predictors. We do this by defining inclusion indicators $\boldsymbol{\gamma}_j^m = (\gamma_{j_1}^m, \ldots, \gamma_{j_{j-1}}^m)$ where $\gamma_{j_k}^m = 1$, indicates that x_{ik} should be included in the regression model for x_{ij} and $\gamma_{j_k}^m = 0$ otherwise. Thus the joint distribution of the predictors in model $\boldsymbol{\gamma}^m$ is

$$p(\boldsymbol{X}|\boldsymbol{\theta}_{\boldsymbol{\gamma}^m}, \boldsymbol{\kappa}) \quad = \quad p(\boldsymbol{x}_1|\theta_1, \kappa_1) \prod_{j=2}^{p} p(\boldsymbol{x}_j|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{\theta}_{j\boldsymbol{\gamma}_j^m}, \kappa_j) \tag{4}$$

where $\boldsymbol{\theta}_{j\boldsymbol{\gamma}_j^m} = (\theta_{j\gamma_j^m 1}, \ldots, \theta_{j\gamma_j^m p_{\gamma_j^m}})'$ are the coefficients in the regression model for $\boldsymbol{x}_j$ in model $\boldsymbol{\gamma}_j^m$ with $p_{\gamma_j^m} = \sum_{k=1}^{j-1} \gamma_{j_k}^m$, $\boldsymbol{\gamma}^m = (\boldsymbol{\gamma}_1^{m'}, \ldots, \boldsymbol{\gamma}_p^{m'})' \in \boldsymbol{\Gamma}^m$ indexes the model characterizing the joint distribution of the predictors, and $\boldsymbol{\theta}_{\boldsymbol{\gamma}^m} = (\theta_1, \boldsymbol{\theta}'_{2\boldsymbol{\gamma}_2^m}, \ldots, \boldsymbol{\theta}'_{p\boldsymbol{\gamma}_p^m})'$.

Thus, the distribution of $\boldsymbol{x}_j$ is conditional on a subset of the predictors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1})$ defined by the inclusion indicators in $\boldsymbol{\gamma}_j^m$. Dropping a predictor $\boldsymbol{x}_k, 1 \leq k \leq j-1$ from the regression model for $\boldsymbol{x}_j$ implies independence between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ conditional on the other predictors in the model and so we are able to incorporate parsimonious relationships between predictors. In the special case when $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j)$ have a multivariate normal distribution this corresponds to putting zeroes in the $(j,k)^{th}$ and $(k,j)^{th}$ entries of its precision matrix.

Therefore, we are performing variable selection on two levels, (1) in the top level model relating the response to predictors, and (2) in the model characterizing the joint distribution of the predictors. In Section 4, we present a theoretical argument implying improved predictive performance for two level variable selection compared with a one level approach that bypasses level (2) and assumes a particular choice of model $\boldsymbol{\gamma}_0^m$ nested in $\boldsymbol{\Gamma}^m$. Note that in the two-level case, there are $2^{\frac{p(p+1)}{2}}$ possible models in the joint model space, $\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}^m$. Hence, even for modest p, the number is enormous. In the next section, we propose a two-level

SSVS algorithm, SSVS^2, which extends the one-level algorithm, SSVS^1.

3 Stochastic Search Variable Selection

The SSVS^2 algorithm described in this section focuses on the case in which $p(\boldsymbol{x}_j \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{\theta}_{\boldsymbol{\gamma}_j^m}, \kappa_j)$ is a normal linear regression model for continuous $\boldsymbol{x}_j$ and is a probit regression model for categorical $\boldsymbol{x}_j$. We also assume a normal or probit form for $p(\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau)$. These special cases are convenient in facilitating use of data augmentation, as proposed by Albert and Chib (1993), to obtain closed forms for conditional model probabilities and posterior distributions. However, the algorithm described can be trivially modified to allow other GLMs through the use of a Laplace approximation to marginal likelihoods used in calculating conditional model probabilities, with adaptive rejection sampling used for updating the parameters from their full conditional posterior distributions given the model.

As for previously-proposed SSVS^1 algorithms, the goal of SSVS^2 is to simultaneously accomplish several goals through the use of an MCMC algorithm that alternates between updating the model indicators and the parameters within the current model. By sampling from full conditional posterior distributions sequentially, the samples converge in distribution to a stationary distribution that is the joint posterior distribution of the model indicators and the parameters within each model. For enormous model spaces, such as the ones encountered in the two-level variable selection problem or the one-level case for moderate to large numbers of candidate predictors, it is not realistic to expect accurate estimates of the exact posterior model probabilities and posterior distributions based on the number of samples it is feasible to collect. Nonetheless, it has been observed that marginal posterior densities of the coefficients for each predictor, marginal inclusion probabilities and predictive distributions tend to be well estimated by SSVS algorithms even in challenging cases.

In Section 3.1 we complete a Bayesian specification of the model with explicit models for each component of the likelihood and with prior distributions for the parameters and model indicators. In Section 3.2 we outline the steps involved in the SSVS^2 algorithm.

3.1 Model and prior specification

We first model the top level which relates the p predictors $\boldsymbol{x}_i$ to the response y_i for each individual i under model $\boldsymbol{\gamma}$. As we are considering y_i to be either continuous or categorical define $y_i = g_y(y_i^*, \boldsymbol{\xi}_y)$, where

$$p(y_i^*|x_{i1}, \ldots, x_{ip}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau) \quad = \quad N(y_i^*; \beta_0 + \boldsymbol{x}'_{\boldsymbol{\gamma}i}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau), \tag{5}$$

$\boldsymbol{x}_{\boldsymbol{\gamma}i}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ as defined in section 2 and τ is the residual variance, with $\phi = \tau^{-1}$. When y_i is continuous g_y is the identity so that $y_i = g_y(y_i^*) = y_i^*$. With an ordered categorical response with $y_i \in \{1, \ldots, c_y\}$, we set $\tau = 1$ and define $\boldsymbol{\xi}_y = (\xi_{y0}, \xi_{y1} \ldots, \xi_{yc_y})'$ to represent the threshold parameters in a generalized probit model with $\xi_{y0} = -\infty, \xi_{yc_y} = \infty$, $\xi_{y1} = 0$ and $y_i = g_y(y_i^*, \boldsymbol{\xi}_y) = \sum_{k=1}^{c_y} kI(\xi_{y,k-1} < y_i^* \leq \xi_{yk})$. In the special case that $c_y = 2$ $y_i = I(y_i^* > 0)$.

We embed all the models within one full model that includes main effects for all the predictors. To simultaneously specify a prior over the model space and for the coefficients within each model, we let

$$\beta_j \quad \sim \quad (1 - \pi_j)\delta_0 + \pi_j N(0; \phi_{\beta_j}^{-1}), \quad \phi_{\beta_j} \sim Ga(1/2, 1/2), \tag{6}$$

where δ_0 is a unit probability mass at zero, $\gamma_j = 1(\beta_j \neq 0)$, and π_j is the prior probability of including the jth predictor, with $\pi_j = 0.5$ if inclusion and exclusion are equally likely. Predictors having zero coefficients are effectively excluded from the model, while a heavy-tailed Cauchy prior is induced through a scale mixture of Gaussians for the coefficients for the included predictors. We place a Jeffreys prior on ϕ for a continuous response and a uniform improper prior for $\boldsymbol{\xi}_y^* = (\xi_{y2}, \ldots, \xi_{y,c_y-1})'$ on the restricted space $\Omega = \{\boldsymbol{\xi}_y^* : 0 < \xi_{y2} < \xi_{y3} < \ldots < \xi_{y,c_y-1}\} \subset R^{c_y-2}$ for a ordered categorical response.

Focusing now on the predictor component model and using the specification of Section 2.3, we let $x_{ij} = g_{x_j}(x_{ij}^*, \boldsymbol{\xi}_{x_j})$ where,

$$p(x_{ij}^* | x_{i1} \ldots x_{i,j-1}, \boldsymbol{\theta}_j, \kappa_j, \boldsymbol{\gamma}_j^m) \quad = \quad N\left(x_{ij}^*; \theta_{j0} + \boldsymbol{x}^{*'}_{j\boldsymbol{\gamma}_j^m i} \boldsymbol{\theta}_{j\boldsymbol{\gamma}_j^m}, \kappa_j\right), \tag{7}$$

where $\boldsymbol{x}^*_{j\boldsymbol{\gamma}_j^m i} = (x^*_{j\gamma_j^m i1}, \ldots, x^*_{j\gamma_j^m ip_{\gamma_j^m}})'$ are the predictors in model $\boldsymbol{\gamma}_j^m$, $\boldsymbol{\theta}_{j\boldsymbol{\gamma}_j^m}$ are the coefficients for these predictors, and $\psi_j = \kappa_j^{-1}$. For continuous x_{ij}, g_{x_j} is the identity so $x_{ij} = g_{x_j}(x_{ij}^*) = x_{ij}^*$ and for an ordered categorical predictor $x_{ij} \in \{1, \ldots, c_{x_j}\}$ set $\psi_j = 1$ and use threshold parameters $\boldsymbol{\xi}_{x_j} = (\xi_{x_j 0}, \xi_{x_j 1} \ldots, \xi_{x_j c_{x_j}})'$ to model $x_{ij} = g_{x_j}(x_{ij}^*, \boldsymbol{\xi}_{x_j}) = \sum_{k=1}^{c_{x_j}} kI(\xi_{x_j,k-1} < x_{ij}^* \leq \xi_{x_j k})$ where, $\xi_{x_j 0} = -\infty$, $\xi_{x_j c_{x_j}} = \infty$ and $\xi_{x_j 1} = 0$. For binary predictors, we let $x_{ij} = I(x_{ij}^* > 0)$.

Define $\boldsymbol{\kappa} = \{\kappa_j, j : x_{ij} = x_{ij}^*\}$ to be the set of scale parameters in the joint distribution of the predictors and $\boldsymbol{X^*} = \{x_{ij}^*, i = 1, \ldots, n,\ j : x_{ij} \neq x_{ij}^*\}$ to be the set of latent variables corresponding to categorical predictors in our data set. To complete a prior specification using a similar specification to (6), we let

$$\theta_{jk} \sim (1 - \pi_{j_k})\delta_0 + \pi_{j_k} N(0, \phi_{\theta_{jk}}^{-1}), \quad \phi_{\theta_{jk}} \sim Ga(1/2, 1/2), \tag{8}$$

for $j = 1, \ldots, p, k = 0, \ldots, j-1$, where $\pi_{j_k} = 0.5$ as a default, $\gamma_{j_k}^m = 1(\theta_{jk} \neq 0)$, and $\boldsymbol{\phi}_{\boldsymbol{\theta}_{\gamma^m}} = \{\phi_{\theta_{jk}}, (j,k) : \gamma_{j_k} = 1\}$. In the SSVS^1 approach we do not perform SSVS on the missing data model, instead we put Jeffreys priors on all regression coefficients and intercepts so that $p(\theta_{j_k}) \propto 1\ j = 1, \ldots, p,\ k = 0, \ldots, j-1$, this implicitly assumes that $\gamma_{j_k}^m = 1$ for all j, k. In both approaches SSVS^2 and SSVS^1 we can again place Jeffreys priors for any residual variances in the regression models and improper uniform priors on the restricted support of the threshold parameters for each categorical predictor.

3.2 Posterior computation

We now outline the basic steps of the SSVS^2 algorithm, focusing for simplicity on the case in which the response is binary and the predictors are binary or continuous. SSVS^2 proceeds by sampling from the joint posterior of the model space $(\boldsymbol{\gamma}, \boldsymbol{\gamma}^m)$, parameters within each model $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\phi}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}^m}, \boldsymbol{\phi}_{\boldsymbol{\theta}_{\boldsymbol{\gamma}^m}}, \boldsymbol{\kappa})$, and the latent variables $(\boldsymbol{y}^*, \boldsymbol{X}^*, \boldsymbol{X}_{mis})$ conditional on the observed data $(\boldsymbol{y}, \boldsymbol{X}_{obs})$.

Under the likelihood and prior specification of Section 3.1, full conditional posterior distributions of each unknown have a simple form allowing Gibbs sampling. These full conditionals are provided in an appendix, and we focus here on updating of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The full conditional posterior of β_j can be expressed as

$$(1-\widehat{\pi}_j)\delta_0 + \widehat{\pi}_j N(E_j, V_j), \tag{9}$$

where $\widehat{\pi}_j$ is the conditional posterior probability of $\gamma_j = 1$, which is

$$\widehat{\pi}_j = 1 - \frac{1-\pi_j}{1-\pi_j+\pi_j \frac{\sqrt{\phi_{\beta_j}}\phi(0)}{V_j^{-\frac{1}{2}}\phi(V_j^{-\frac{1}{2}}E_j)}},$$

and the conditional expectation and variance of β_j given $\gamma_j = 1$ are

$$E_j = V_j \sum_{i=1}^{n} x_{ij}\tilde{y}^*_{i_j}, \quad V_j = \left(\phi_{\beta_j} + \sum_{i=1}^{n} x_{ij}^2\right)^{-1},$$

with $\tilde{y}^*_{i_j} = y^*_i - \beta_0 - \sum_{h \neq j} x_{ih}\beta_h$ and $\phi(.)$ the standard normal density.

Note that in updating β_j, we automatically update $\gamma_j = 1(\beta_j \neq 0)$. Upon convergence, samples of $\boldsymbol{\gamma}$ are drawn from the marginal posterior distribution $p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}_{obs})$. A model's posterior probability can then be estimated by the proportion of samples in that model. In addition, marginal inclusion probabilities, $\Pr(\gamma_j = 1 \,|\, \boldsymbol{y}, \boldsymbol{X}_{obs})$, provide a convenient weight of evidence that the jth predictor should be included.

The full conditional posterior of θ_{jk} is

$$(1-\widehat{\pi}_{j_k})\delta_0 + \widehat{\pi}_{j_k} N(E_{jk}, V_{jk}), \tag{10}$$

where the conditional posterior probability of $\gamma_{jk}^m = 1$ is

$$\widehat{\pi}_{j_k} = 1 - \frac{1-\pi_{j_k}}{1-\pi_{j_k} + \pi_{j_k}\frac{\sqrt{\phi_{\theta_{jk}}}\phi(0)}{V_{jk}^{-\frac{1}{2}}\phi(V_{jk}^{-\frac{1}{2}}E_{jk})}},$$

and the conditional posterior mean and variance given inclusion is

$$E_{jk} = V_{jk}\psi_j \sum_{i=1}^{n} x_{ik}^{*}\tilde{x}_{ij_k}, \quad V_{jk} = \left(\phi_{\theta_{jk}} + \psi_j \sum_{i=1}^{n} x_{ik}^{*^2}\right)^{-1},$$

with $\tilde{x}_{ij_k} = x_{ij}^{*} - \theta_{j0} - \sum_{h=1,h\neq k}^{j-1} x_{ih}^{*}\theta_{jh}$. All other parameters can be sampled from their full conditionals as standard in regression models. For details of all the full conditionals to implement the Gibbs sampler please refer to the Appendix.

The missing predictors are also imputed from their full conditional distributions, which are available in closed form, and so we embed the imputation of missing covariates within our stochastic search of the model space, allowing simultaneous treatment of the missing data and variable selection problems. We evaluate both SSVS^1 and SSVS^2 by considering posterior model inferences as well as out of sample predictive performance in a simulation study. We compare our results to model averaging performed on the original completely observed data (prior to introducing covariate missingness). More details on this are presented in Section 5. In the next section we provide a theoretical argument supporting the use of the SSVS^2 approach.

4 Improved Predictive Performance

Raftery *et al.* (1997) showed that Bayesian model averaging has better predictive performance than using any single model alone. We present a similar argument here extending this to the case of model averaging over the missing data models as opposed to using a single model for imputations.

Let Δ be the quantity we are interested in predicting (e.g. the disease outcome of a patient). Denote $(\boldsymbol{y}, \boldsymbol{X}_{obs})$ to represent the observed data and $\boldsymbol{X}_{mis}$ the missing data. As before let $\boldsymbol{\gamma}$ index the set $\boldsymbol{\Gamma}$ of the 2^p possible models for the response $\boldsymbol{y}$ and let $\boldsymbol{\gamma}^m$ index the set $\boldsymbol{\Gamma}^m$ of the $2^{\frac{p(p-1)}{2}}$ models for the missing data $\boldsymbol{X}_{mis}$, with γ_0^m representing a particular choice of model nested in $\boldsymbol{\Gamma}^m$. Then in our SSVS^2 approach we define the predictive distribution of Δ as f, where

$$f = \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \sum_{\boldsymbol{\gamma}^m\in\boldsymbol{\Gamma}^m} \left[\int p(\Delta|\boldsymbol{\gamma}, \boldsymbol{\gamma}^m, \boldsymbol{y}, \boldsymbol{X}_{obs}, \boldsymbol{X}_{mis}) p(\boldsymbol{X}_{mis}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^m, y, \boldsymbol{X}_{obs}) d\boldsymbol{X}_{mis} \right] p(\boldsymbol{\gamma}, \boldsymbol{\gamma}^m|\boldsymbol{y}, \boldsymbol{X}_{obs})$$

and in a one level approach the predictive distribution of Δ, defined as g is,

$$g = \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \left[\int p(\Delta|\boldsymbol{\gamma}, \boldsymbol{\gamma}_0^m, \boldsymbol{y}, \boldsymbol{X}_{obs}, \boldsymbol{X}_{mis}) p(\boldsymbol{X}_{mis}|\boldsymbol{\gamma}, \boldsymbol{\gamma}_0^m, \boldsymbol{y}, \boldsymbol{X}_{obs}) d\boldsymbol{X}_{mis} \right] p(\boldsymbol{\gamma}|\boldsymbol{\gamma}_0^m, \boldsymbol{y}, \boldsymbol{X}_{obs})$$

The following theorem then holds:

Theorem 4.1. $-E_f[log(f(x)] \; \leq \; -E_f[log(g(x))]$

Hence under a logarithmic scoring rule we see that the estimate that model averages over both the set of models for the response as well as the missing data has a lower risk than the one which does not model average over the missing data.

Proof. Consider the Kullback-Leibler divergence for f and g:

$K(f:g) = \int_{-\infty}^{\infty} f(x) log\left(\frac{f(x)}{g(x)}\right) dx \quad \geq \quad 0$ (by non-negativity of K-L divergence).

$\Rightarrow \int_{-\infty}^{\infty} \{f(x) log\,(f(x)) - f(x) log\,(g(x))\} dx \quad \geq \quad 0$

$\Rightarrow \int_{-\infty}^{\infty} f(x) log\,(f(x))\, dx \quad \geq \quad \int_{-\infty}^{\infty} f(x) log\,(g(x))\, dx$

$\Rightarrow E_f[log(f(x)] \;\; \geq \;\; E_f[log(g(x))]$

$\Rightarrow -E_f[log(f(x)] \;\; \leq \;\; -E_f[log(g(x))]$

□

We can use the above theorem to justify our SSVS^2 approach having better predictive performance for a future units' observation y^{new} conditional on observing some of its covariates $\boldsymbol{x}^{new}$ with $\boldsymbol{m}^{new}$ indicating which elements in $\boldsymbol{x}^{new}$ are missing. Defining $\boldsymbol{x}^{new}_{obs} = \{x^{new}_j, j : m^{new}_j = 0\}$ we express the predictive density of the future observation y^{new} in the SSVS^2 approach as:

$$p(y^{new}|\boldsymbol{x}^{new}_{obs}) = \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \sum_{\boldsymbol{\gamma}^m\in\boldsymbol{\Gamma}^m} p(y^{new}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^m, \boldsymbol{x}^{new}_{obs}) p(\boldsymbol{\gamma}, \boldsymbol{\gamma}^m|\boldsymbol{y}, \boldsymbol{X}_{obs})$$

which we know, by using the result of Theorem 4.1, has better predictive properties than restricting imputation models to be based on using a single model $\boldsymbol{\gamma}^m_0$, where

$$p(y^{new}|\boldsymbol{x}^{new}_{obs}, \boldsymbol{\gamma}^m_0) = \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} p(y^{new}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^m_0, \boldsymbol{x}^{new}_{obs}) p(\boldsymbol{\gamma}|\boldsymbol{\gamma}^m_0, \boldsymbol{y}, \boldsymbol{X}_{obs})$$

.

Note that for Theorem 4.1 to apply we must have model $\boldsymbol{\gamma}^m_0$ nested within the set of models being considered in the SSVS^2 approach, $\boldsymbol{\Gamma}^m$. This is not the case with the

SSVS^1 approach described in Section 3.1 as we are using Jeffreys priors, not Cauchy priors, for the regression coefficients in the predictor component of the model. Nevertheless, we might still expect improved predictive performance when using the SSVS^2 approach, as the SSVS^1 approach does not incorporate model uncertainty into the joint distribution of the predictors. In the next section we investigate possible gains in predictive performance as well as additional benefits in obtaining posterior inferences of the SSVS^2 approach in a simulation study.

5 Simulation Studies

We simulate 1000 units with 17 predictors and a binary response y using a probit model. Approximately half the units' responses were assigned to either 0 or 1. Of the 17 predictors only 4 were used to generate the response variable and we denote these to be our true predictors, the rest we denote as null predictors. Any relationship between the null predictors and the response is spurious and is due to canonical correlations with the true set of predictors. In addition, we specify sparse relationships between the predictors using a DAG set up where $\boldsymbol{x}_j$ is simulated conditional on a subset of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1})$. Half of the observations are assigned to be in our training data set and we use the other half as an out of sample test data set.

To evaluate our two approaches (SSVS^2 and SSVS^1) we introduce missing values in the covariates, where we use relationships similar to those used to simulate the data to generate the missing values. Each predictor is set to have approximately 40% of its values missing. We can then perform variable selection via the Gibbs sampler outlined in section 3 using both approaches and compare posterior model inferences. In addition we can consider posterior inferences in the situation when there is no missing data (SSVS^{obs}). Figure 1 presents the mean inclusion and exclusion probabilities of the true and null predictor sets respectively across different training data sizes.

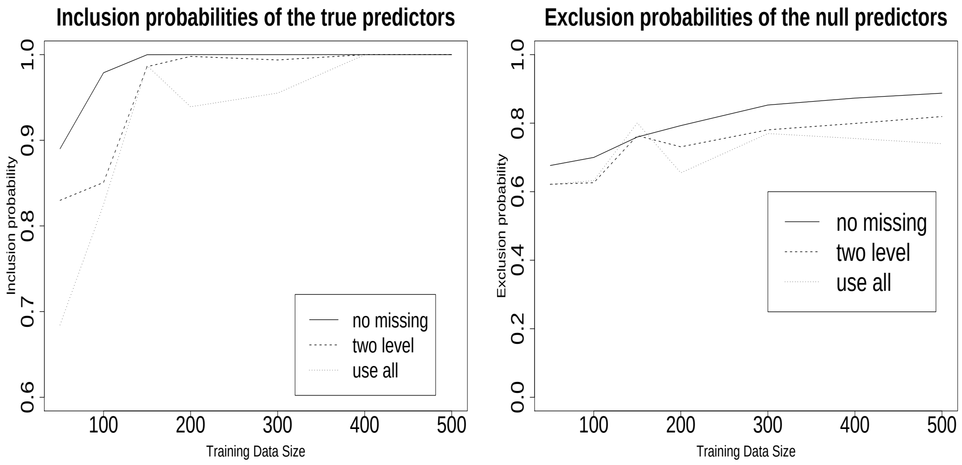

Figure 1: Mean Inclusion/Exclusion Probabilities for the True/Null Predictors respectively for the three cases across different training data sizes

The closer the line is to 1 in both plots the better the method is performing. As expected the case of fully observed covariate information does the best with performance increasing with training data size. In the first plot the SSVS^2 approach also exhibits a similar monotone pattern with gains in estimation of the true predictors' inclusion probabilities over SSVS^1 evident. In the second plot there is not much difference between the two approaches, with small gains in estimation of the null predictors' exclusion probabilities as the training data size increases.

In addition we can use the out of sample test data set to evaluate predictive performance of the methods. We impute missing covariate values in the test data set from their full conditional distributions within each iteration of the MCMC (see the Appendix for details) and can thus generate predictions for the response which can be compared with the actual values. As we have a binary response this can be conveniently summarized by the percentage of units correctly classified. We present a plot of the correct classification rate for the two approaches plus the situation when there is no missing data against training data size in figure 2.

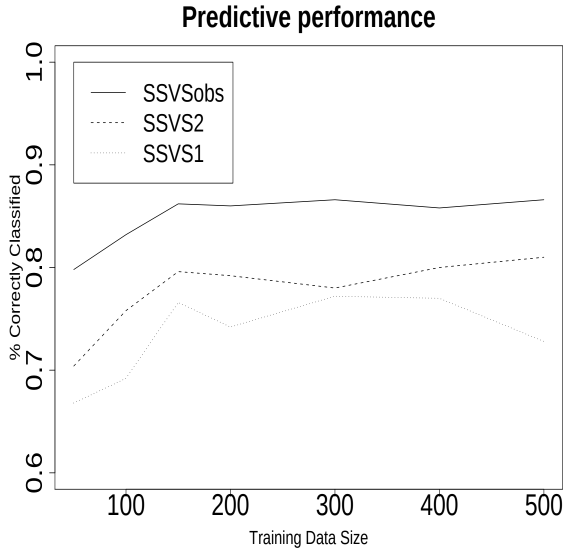

Figure 2: Out of sample predictive performance for the two different methods compared to the case with fully observed covariates

We see that of course the situation when the covariates are all fully observed has the best classification rates. The SSVS^2 method does better than the SSVS^1 approach across all training data sizes. In SSVS^{obs} and SSVS^2 the correct classification rate tends to increase with training data size, while the increasing trend is not so clear with SSVS^1.

6 Reproductive Epidemiology Application

We now apply our methods to data from the Longnecker *et al.* (2001) sub-study of the US Collaborative Perinatal Project (CPP). We are interested in predicting high risk pregnancies for women with advanced maternal age (35 or older) when there are missing predictors, for related work refer to (Eastaugh *et al.*, 1997). We took our response to be whether a preterm birth was observed or not and chose thirteen fully observed variables (binary and continuous) as candidate predictors. We include mother's age, height, pre-pregnancy BMI, pregnancy weight gain, smoking status, race, serum total cholesterol, triglycerides, sum of PCBs, and p,p'-DDE (lipid adjusted). We also include the child's gender, the socio-economic index and whether the prenatal care was adequate. The sample size was 182.

We then introduced approximately 40% missing data in each predictor. In particular we generate missing values in the predictors race, pre-pregnancy BMI and socio-economic index using an underlying latent lifestyle factor that we assume is related to these three predictors and the response preterm birth. For all other predictors we generate missing values using an MCAR mechanism.

We evaluate the performance of the SSVS^2 and SSVS^1 approaches by comparing the posterior means of each regression coefficient to the posterior means obtained from SSVS^{obs}. Figure 3 plots the absolute value differences in posterior means obtained from SSVS^1 to the posterior means obtained from SSVS^{obs} for each regression coefficient against similar absolute value differences when using SSVS^2. Points above the line $y = x$ (included on the plot) indicate better performance in SSVS^2 over SSVS^1 and vice versa. We see that there are several points quite far above the line and so there is some evidence to suggest that SSVS^2 is doing better than SSVS^1 in obtaining closer estimates of the posterior mean to those obtained using SSVS^{obs}.

7 Conclusion

In this paper we presented an efficient way to accommodate the problem of missing continuous and binary covariates when model averaging in Generalized Linear Models. We illustrated the benefits of additionally model averaging over the imputation models in posterior inferences and out of sample predictive performance through a simulation study. Finally, we applied our method to a reproductive epidemiology study to evaluate the benefits in using our two level approach. We found that SSVS^2 obtained estimates closer to the estimates

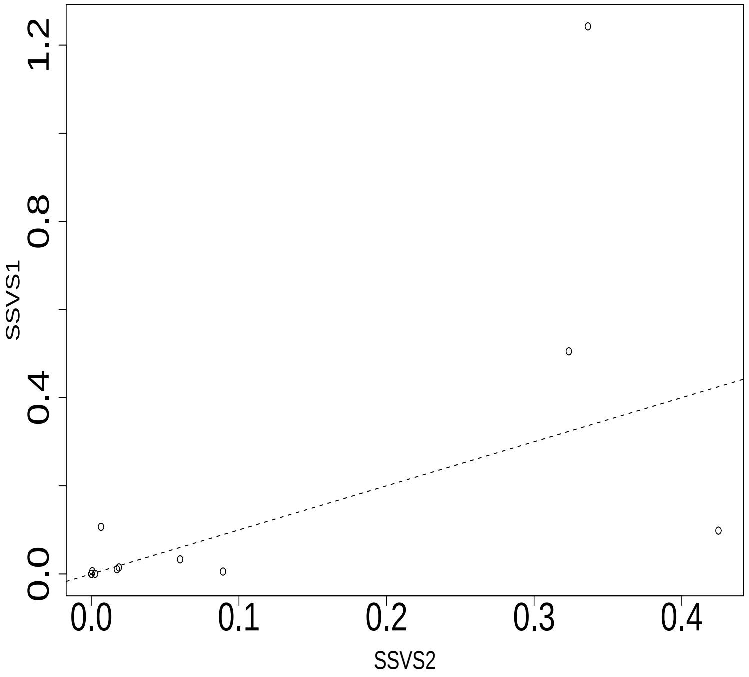

Figure 3: Absolute difference in posterior means of regression coefficients from SSVS^1 against SSVS^2 as compared to SSVS^{obs}, line $y = x$ included

from SSVS^{obs} than those from SSVS^1.

It would be interesting to extend our models to incorporate a wider range of Generalized Linear Models such as count response data, perhaps using approximations to the Marginal Likelihood developed by Raftery (1996) or Cai and Dunson (2006). We could also in principle extend our method to mixed effects data where variable selection could be performed on both the fixed effects regression coefficients as well as the variances of the random effects (Kinney and Dunson, 2007). An alternative prior specification that takes into account the scale of the predictors such as Zellner's g prior might also be preferable to the ridge type priors used in this paper.

References

Cai, B. and Dunson, D. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics* **62**, 446–457.

Clyde, M. and George, E. (2004). Model uncertainty. *Statistical Science* **19**, 81–94.

Eastaugh, J., Smye, S., Snowden, S., Walker, J., Dear, P., and Farrin, A. (1997). Comparison of neural networks and statistical models to predict gestational age at birth. *Neural Computing & Applications* **6**, 158–164.

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

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

Ibrahim, J., Lipsitz, S., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Ser. B* **61**, 173–190.

Kinney, S. K. and Dunson, D. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics* **63**, 690–698.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association* **103**, 410–423.

Lipsitz, S. and Ibrahim, J. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika* **83**, 916–922.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. New York: John Wiley & Sons.

Longnecker, M. P., Klebanoff, M. A., Zhou, H., and Brock, J. W. (2001). Association between maternal serum concentration of the ddt metabolite dde and preterm and small-for-gestational-age babies at birth. *Lancet* **358**, 110–114.

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

Raftery, A. (1996). Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 251–266.

Raftery, A., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179–191.

Yang, X., Bellin, T., and Boscardin, W. (2006). Imputation and variable selection in linear regression models with missing covariates. *Biometrics* **61**, 498–506.

8 Appendix - Full Conditionals

We present here the joint posterior distribution and the resulting full conditionals required for the Gibbs sampler in the SSVS^2 approach, focusing for simplicity on the case in which the response is binary and the predictors are binary or continuous.

SSVS proceeds by sampling from the joint posterior of the model space $(\boldsymbol{\gamma}, \boldsymbol{\gamma^m})$, parameters within each model $(\boldsymbol{\beta_\gamma}, \boldsymbol{\phi_{\beta_\gamma}}, \boldsymbol{\theta_{\gamma^m}}, \boldsymbol{\phi_{\theta_{\gamma^m}}}, \boldsymbol{\kappa})$, and the latent/unobserved variables $(\boldsymbol{y^*}, \boldsymbol{X^*}, \boldsymbol{X}_{mis})$ conditional on the observed data $(\boldsymbol{y}, \boldsymbol{X}_{obs})$. The joint posterior is expressed below,

$$
\begin{aligned}
&\pi(\boldsymbol{\gamma, \gamma^m, y^*}, \boldsymbol{\beta_\gamma}, \boldsymbol{\phi_{\beta_\gamma}}, \boldsymbol{\theta_{\gamma^m}}, \boldsymbol{\phi_{\theta_{\gamma^m}}}, \boldsymbol{\kappa}, \boldsymbol{X}^*, \boldsymbol{X}_{mis} | \boldsymbol{y}, \boldsymbol{X}_{obs}) \\
&\propto \left\{ \textstyle\prod_i^n p(y_i|y_i^*) p(y_i^*|\boldsymbol{\beta_\gamma}, \boldsymbol{x_i}) p(\boldsymbol{x_i, x_i^*}|\boldsymbol{\theta_{\gamma^m}}, \boldsymbol{\kappa}) \right\} p(\boldsymbol{\beta_\gamma}, \boldsymbol{\phi_{\beta_\gamma}}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \pi(\boldsymbol{\theta_{\gamma^m}}, \boldsymbol{\phi_{\theta_{\gamma^m}}}|\boldsymbol{\gamma^m}) p(\boldsymbol{\gamma^m}) p(\boldsymbol{\kappa}) \\
&\propto \left\{ \textstyle\prod_i^n \left(y_i I(y_i^* \geq 0) + (1 - y_i) I(y_i^* < 0) \right) N(y_i^*; \beta_0 + \boldsymbol{x'_{\gamma i}} \boldsymbol{\beta_\gamma}, 1) \right. \\
&\quad \times \left. \left[\textstyle\prod_{j=1}^p p(x_{ij}|x_{ij}^*) N \left(x_{ij}^*; \theta_{j0} + \boldsymbol{x^{*\prime}_{j\gamma_j^m i}} \boldsymbol{\theta_{j\gamma_j^m}}, \kappa_j \right) \right] \right\} \\
&\quad \times \left\{ \textstyle\prod_{j=0}^p p(\beta_j|\phi_{\beta_j}, \gamma_j) p(\gamma_j) p(\phi_{\beta_j}|\gamma_j) \right\} \left\{ \textstyle\prod_{j=1}^p \prod_{k=0}^{j-1} \left[p(\theta_{jk}|\phi_{\theta_{jk}}, \gamma_{j_k}^m) p(\phi_{\theta_{jk}}|\gamma_{j_k}^m) p(\gamma_{j_k}^m) \right] \right\} p(\boldsymbol{\kappa})
\end{aligned}
$$

where, $p(x_{ij}|x_{ij}^*) = x_{ij} I(x_{ij}^* \geq 0) + (1 - x_{ij}) I(x_{ij}^* < 0)$, $\kappa_j = 1$ for binary x_{ij} and $p(x_{ij}|x_{ij}^*) = \delta_{x_{ij}^*}(x_{ij})$ for continuous x_{ij}

With the models for the data and the prior distributions for the parameters discussed in section 3, the full conditionals are available in closed form. First consider the full conditional distributions for the parameters in the predictor component of the model. Sample θ_{jk} from,

$$
(1 - \widehat{\pi}_{j_k}) \delta_0 + \widehat{\pi}_{j_k} N(E_{jk}, V_{jk}), \tag{11}
$$

where the conditional posterior probability of $\gamma_{jk}^m = 1$ is

$$\widehat{\pi}_{j_k} = 1 - \frac{1 - p_{j_k}}{1 - p_{j_k} + p_{j_k} \frac{\sqrt{\phi_{\theta_{jk}}}\phi(0)}{V_{jk}^{-\frac{1}{2}} \phi(V_{jk}^{-\frac{1}{2}} E_{jk})}},$$

and the conditional posterior mean and variance given inclusion is

$$E_{jk} = V_{jk}\psi_j \sum_{i=1}^{n} x_{ik}^* \tilde{x}_{ij_k}^*, \quad V_{jk} = \left(\phi_{\theta_{jk}} + \psi_j \sum_{i=1}^{n} x_{ik}^{*^2} \right)^{-1},$$

with $\tilde{x}_{ij_k}^* = x_{ij}^* - \theta_{j0} - \sum_{h=1,h\neq k}^{j-1} x_{ih}^* \theta_{jh}$ and $\phi(.)$ the standard normal density. Also update $\phi_{\theta_{jk}}$ for predictors included in the model from,

$$Ga\left(1, \frac{\theta_{jk}^2 + 1}{2}\right). \tag{12}$$

Next sample $\psi_k = \kappa_k^{-1}$ for continuous $\boldsymbol{x}_k$ from,

$$Ga\left(\frac{n}{2}\ ,\ \left(\sum_{i=1}^{n} x_{ik}^* - (\theta_{j0} - \sum_{j=1}^{k} x_{ij}^* \theta_{jk})\right)^2\right) \tag{13}$$

while $\psi_k = \kappa_k^{-1} = 1$ for binary $\boldsymbol{x}_k$. Now for the i^{th} missing continuous covariate value x_{ij}, we impute from a normal distribution,

$$N\left(\tilde{\psi}_j^{-1}\tilde{\mu}_{ij}, \tilde{\psi}_j^{-1}\right) \tag{14}$$

where,

$$\begin{aligned}
\tilde{\psi}_j &= \beta_j^2 + \psi_j + \sum_{k=j+1}^{p} \psi_j \theta_{kj}^2, \\
\tilde{\mu}_{ij} &= \tilde{y}_{i_j}^* \beta_j + \psi_j \mu_{ij} + \sum_{k=j+1}^{p} \psi_k \theta_{kj} \tilde{x}_{ik_j}^*
\end{aligned}$$

and,

$$\begin{aligned}
\tilde{y}_{i_j}^* &= y_i^* - \beta_0 - \sum_{h \neq j} x_{ih} \beta_h, \\
\mu_{ij} &= \theta_{j0} + \sum_{k=1}^{j-1} x_{ik}^* \theta_{jk}, \\
\tilde{x}_{ik_j}^* &= x_{ik}^* - \theta_{k0} - \sum_{h=1, h \neq j}^{k-1} x_{ih}^* \theta_{kh}.
\end{aligned}$$

While when x_{ij} is binary and missing, we first impute its underlying latent variable x_{ij}^* from the full conditional,

$$\tilde{\pi}_{ij} N_+ \left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1} \right) + (1 - \tilde{\pi}_i) N_- \left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1} \right) \tag{15}$$

where,

$$\begin{aligned}
\tilde{\psi}_j^{-1} &= \psi_j + \sum_{k=j+1}^{p} \psi_j \theta_{kj}^2, \\
\tilde{\mu}_{ij} &= \psi_j \mu_{ij} + \sum_{k=j+1}^{p} \psi_k \theta_{kj} \tilde{x}_{ik_j}^*, \\
\tilde{\pi}_{ij} &= \frac{(1 - \Phi\left(\frac{\tilde{\mu}_{ij}}{\sqrt{\tilde{\psi}_j}}\right)) \phi\left(\tilde{y}_{i_j}^* - \beta_j\right)}{(1 - \Phi\left(\frac{\tilde{\mu}_{ij}}{\sqrt{\tilde{\psi}_j}}\right)) \phi\left(\tilde{y}_{i_j}^* - \beta_j\right) + \Phi\left(\frac{\tilde{\mu}_{ij}}{\sqrt{\tilde{\psi}_j}}\right) \phi\left(\tilde{y}_{i_j}^*\right)}
\end{aligned}$$

and,

$$\begin{aligned}\mu_{ij} &= \theta_{j0} + \sum_{k=1}^{j-1} x^*_{ik}\theta_{jk}, \\ \tilde{x}^*_{ik_j} &= x^*_{ik} - \theta_{k0} - \sum_{h=1,h\neq j}^{k-1} x^*_{ih}\theta_{kh}, \\ \tilde{y}^*_{i_j} &= y^*_i - \beta_0 - \sum_{h\neq j} x_{ih}\beta_h,\end{aligned}$$

and then impute $x_{ij} = I(x^*_{ij} > 0)$. We also update latent x^*_{ij} for observed binary x_{ij} from the following distribution:

$$x_{ij} N_+ \left(\tilde{\psi}_j^{-1}\tilde{\mu}_{ij}, \tilde{\psi}_j^{-1}\right) + (1 - x_{ij}) N_- \left(\tilde{\psi}_j^{-1}\tilde{\mu}_{ij}, \tilde{\psi}_j^{-1}\right) \tag{16}$$

where, $\tilde{\psi}$ and $\tilde{\mu}_{ij}$ are as in (15). Note that for individual i predictors other than x_{ij} may be missing, in the imputations we condition on the most recently imputed values of other missing predictors. Now conditional on the observed and imputed predictors we can sample from the full conditionals in the top level models for the response. We sample β_j from its full conditional posterior,

$$(1 - \widehat{\pi}_j)\delta_0 + \widehat{\pi}_j N(E_j, V_j), \tag{17}$$

where $\widehat{\pi}_j$ is the conditional posterior probability of $\gamma_j = 1$, which is

$$\widehat{\pi}_j = 1 - \frac{1 - \pi_j}{1 - \pi_j + \pi_j \frac{\sqrt{\phi_{\beta_j}}\phi(0)}{V_j^{-\frac{1}{2}}\phi(V_j^{-\frac{1}{2}}E_j)}},$$

and the conditional expectation and mean of β_j given $\gamma_j = 1$ are

$$E_j = V_j \sum_{i=1}^{n} x_{ij} \tilde{y}^*_{i_j}, \quad V_j = \left(\phi_{\beta_j} + \sum_{i=1}^{n} x_{ij}^2 \right)^{-1},$$

with $\tilde{y}^*_{i_j} = y^*_i - \beta_0 - \sum_{h \neq j} x_{ih}\beta_h$ and $\phi(.)$ the standard normal density. We sample ϕ_{β_j} for predictors included in the model from its full conditional,

$$Ga\left(1, \frac{\beta_j^2 + 1}{2}\right). \tag{18}$$

Finally sample y^*_i from its full conditional,

$$y_i N_+(\boldsymbol{x_i'\beta}, 1) + (1 - y_i) N_-(\boldsymbol{x_i'\beta}, 1) \tag{19}$$

In this way within one Gibbs sampler we repeatedly impute values for the missing covariates from their full conditional distributions and conditional on the completed data set perform variable selection on the model relating the response to the predictors.

When imputing missing values in the out of sample test data we do not observe the response y and so we must impute from modified full conditionals. For x_{ij} missing and continuous impute from,

$$N\left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1}\right) \tag{20}$$

where, $\tilde{\psi}_j$ and $\tilde{\mu}_{ij}$ are as in (15). For x_{ij} missing and binary impute its underlying latent variable x^*_{ij} from (20) and impute $x_{ij} = I(x^*_{ij} > 0)$. For x_{ij} observed and binary update x^*_{ij} from the same distribution as (16).