AUSTRIAN JOURNAL OF STATISTICS
Volume 40 (2011), Number 4, 241–264

Comparing Spike and Slab Priors for Bayesian Variable Selection

Gertraud Malsiner-Walli and Helga Wagner
Johannes Kepler Universität Linz, Austria

Abstract: An important task in building regression models is to decide which regressors should be included in the final model. In a Bayesian approach, variable selection can be performed using mixture priors with a spike and a slab component for the effects subject to selection. As the spike is concentrated at zero, variable selection is based on the probability of assigning the corresponding regression effect to the slab component. These posterior inclusion probabilities can be determined by MCMC sampling. In this paper we compare the MCMC implementations for several spike and slab priors with regard to posterior inclusion probabilities and their sampling efficiency for simulated data. Further, we investigate posterior inclusion probabilities analytically for different slabs in two simple settings. Application of variable selection with spike and slab priors is illustrated on a data set of psychiatric patients where the goal is to identify covariates affecting metabolism.

Zusammenfassung: Ein wesentliches Problem der Regressionsmodellierung ist die Auswahl der Regressoren, die ins Modell aufgenommen werden. In einem Bayes-Ansatz kann Variablenselektion durchgeführt werden, indem als a-priori-Verteilung für die Regressionseffekte der in Frage kommenden Variablen eine Mischverteilung mit zwei Komponenten gewählt wird: die erste Komponente mit einer Spitze bei Null wird als "spike", die zweite flache Komponente als "slab" bezeichnet. Die Selektion der Variablen erfolgt dann auf Basis der posteriori Wahrscheinlichkeit, mit der ein Effekt der slab-Komponente zugeordnet wird. Diese sogenannten Inklusionswahrscheinlichkeiten können mit Hilfe der MCMC-Ziehungen geschätzt werden. Im vorliegenden Beitrag werden MCMC-Implementierungen für verschiedene spike-and-slab-Verteilungen hinsichtlich der Inklusionswahrscheinlichkeiten und der Effizienz ihrer Schätzung anhand von simulierten Daten verglichen. Außerdem untersuchen wir die Inklusionswahrscheinlichkeiten für verschiedene Slab-Komponenten in zwei einfachen Fällen auch analytisch. Schließlich wird Variablenselektion mit Spike-and-Slab-Priori-Verteilungen auf einen medizinischen Datensatz angewendet, um Regressoren, die den Stoffwechsel von psychiatrischen Patienten beeinflussen, zu indetnifizieren.

Keywords: Dirac Spike, SSVS, NMIG prior, Normal Scale Mixtures, Posterior Inclusion Probability.

1 Introduction

A major task in building a regression model is to select those regressors from a large set of potential covariates which should be included in the final model. Correct classification

of regressors as having (nearly) zero or non-zero effects is important: omitting regressors with non-zero effect will lead to biased estimates whereas inclusion of regressors with zero effect causes loss in estimation precision and predictive performance of the model.

For the regression coefficients, many Bayesian variable selection methods use mixture priors with two components: a spike concentrated around zero and a comparably flat slab. In this paper we compare spike and slab priors with two different specifications for the spike: absolutely continuous and spikes defined by a point mass at zero, so called Dirac spikes. We consider here Dirac spikes combined with different normal slabs and priors where both spike and slab are normal distributions as in George and McCulloch (1993) or scale mixtures of normals as in Ishwaran and Rao (2005) and Konrath, Kneib, and Fahrmeir (2008).

Bayesian variable selection with spike and slab priors can be accomplished by MCMC methods, but depending on the type of the spike the specific implementations differ: A Dirac spike requires computation of marginal likelihoods, i.e. integrating over the parameters subject to selection, in each MCMC iteration. This is not necessary for spikes specified by an absolutely continuous distribution. However, regression effects are not shrunk exactly to zero and therefore the dimension of the model is not reduced during MCMC. In this paper we compare posterior inclusion probabilities under different spike and slab priors as well as their MCMC sampling efficiency.

The rest of the paper is structured as follows. Section 2 describes the basic model and the two types of spike and slab priors. Implementation of MCMC sampling schemes is outlined for both spike types in Section 3 and Section 4 presents results from a simulation study comparing five different spike and slab priors on simulated data. To get further insight, posterior inclusion probabilities are investigated analytically in two simple settings for Dirac spikes combined with different slabs in Section 5. Section 6 illustrates application of Bayesian variable selection on a data set where the goal is to identify covariates which have an effect on metabolism of psychiatric patients. Finally, Section 7 summarizes the results and indicates modifications for the slab component to be considered in further research.

2 Model Specification

2.1 The Linear Regression Model

In the standard linear regression model the outcome $\mathbf{y} = (y_1, \ldots, y_N)$ of subjects $i = 1, \ldots, N$ is modeled as a linear function of the regressors with a Gaussian error term,

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}\,, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma^2\right)\,. \tag{1}$$

Here $\boldsymbol{\alpha}$ is the $d \times 1$ vector of regression coefficients. We assume that the covariate vectors are centered with the null vector as mean, so that $\mathbf{X}'\mathbf{1} = \mathbf{0}$ and the mean μ is constant over all models. As the columns of the design matrix are orthogonal to the unit vector, the log-likelihood can be written as

$$l(\mathbf{y}|\mu, \boldsymbol{\alpha}, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Big(N(\bar{y} - \mu)^2 + (\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})'(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})\Big)\,,$$

where $\mathbf{y}_c = \mathbf{y} - \bar{y}$ denotes the vector of centered responses.

In a Bayesian approach, model specification is completed with priors for the model parameters $(\mu, \sigma^2, \boldsymbol{\alpha})$. We assume a prior of the structure $p(\mu, \sigma^2, \boldsymbol{\alpha}) = p(\mu, \sigma^2)p(\boldsymbol{\alpha}|\sigma^2, \mu)$ with the usual uninformative prior for mean and error variance

$$p(\mu, \sigma^2) = \frac{1}{\sigma^2}, \tag{2}$$

and use spike and slab priors for the regression coefficients $\boldsymbol{\alpha}$.

2.2 Spike and Slab Priors

Mixture priors with spike and slab components have been used extensively for variable selection, see e.g. Mitchell and Beauchamp (1988), George and McCulloch (1993, 1997) and Ishwaran and Rao (2005). The spike component, which concentrates its mass at values close to zero, allows shrinkage of small effects to zero, whereas the slab component has its mass spread over a wide range of plausible values for the regression coefficients. To specify spike and slab priors we introduce indicator variables $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_d)$ where δ_j takes the value 1, if α_j is allocated to the slab component and we denote by $\boldsymbol{\alpha}_{\boldsymbol{\delta}}$ the vector comprising those elements of $\boldsymbol{\alpha}$ where $\delta_j = 1$. We consider priors, where regression effects allocated to the spike component are independent of each other and independent of $\boldsymbol{\alpha}_{\boldsymbol{\delta}}$ a priori, whereas elements of $\boldsymbol{\alpha}_{\boldsymbol{\delta}}$ may be dependent. These spike and slab priors can be written as

$$p(\boldsymbol{\alpha}|\boldsymbol{\delta}) = p_{\text{slab}}(\boldsymbol{\alpha}_{\boldsymbol{\delta}}) \prod_{j:\delta_j=0} p_{\text{spike}}(\alpha_j)\,,$$

where p_{spike} and p_{slab} denote the univariate spike and the multivariate slab distribution respectively. The prior inclusion probability $p(\delta_j = 1)$ of the effect α_j is specified hierarchically as

$$p(\delta_j = 1|\omega) = \omega\,, \qquad \omega \sim \mathcal{B}(a_\omega, b_\omega)\,.$$

Note, that the indicator variables δ_j are independent conditional on the prior inclusion probability ω, but dependent marginally. This is eventually not justified in practical applications and could be relaxed by using an individual inclusion probability ω_j for each regression effect α_j,

$$p(\delta_j = 1|\omega_j) = \omega_j\,, \qquad \omega_j \sim \mathcal{B}(a_{\omega_j}, b_{\omega_j})\,.$$

Prior information on individual inclusion probabilities could be incorporated by appropriate choice of the parameters a_{ω_j} and b_{ω_j}.

The introduction of indicator variables allows classification of regression effects as (practically) zero, if $\delta_j = 0$ and non-zero otherwise. Variable selection is based on the posterior probability of assigning the corresponding regression effect to the slab component, i.e. the posterior inclusion probability $p(\delta_j = 1|\mathbf{y})$, which can be sampled by MCMC methods. Basically two different types of spikes have been proposed in the literature: Spikes specified by an absolutely continuous distribution and spikes specified by a point mass at zero, called Dirac spikes. Specifications of priors for both spike types, which are compared in this paper, are presented in more detail in the following sections.

2.2.1 Absolutely Continuous Spikes

To specify an absolutely continuous spike, in principle any unimodal continuous distribution with mode at zero could be used. Usually absolutely continuous spikes are combined with slabs, where the components of $\boldsymbol{\alpha}_{\boldsymbol{\delta}}$ are independent conditional on $\boldsymbol{\delta}$, i.e.

$$p_{\text{slab}}(\boldsymbol{\alpha}_{\boldsymbol{\delta}}) = \prod_{j:\delta_j=1} p_{\text{slab}}(\alpha_j)\,.$$

Here we consider priors spike and slab components are specified by the same distribution family but with a variance ratio r considerably smaller than 1,

$$r = \frac{\text{var}_{\text{spike}}(\alpha_j)}{\text{var}_{\text{slab}}(\alpha_j)} \ll 1\,. \tag{3}$$

We use only spikes and slabs which can be represented as scale mixtures of normal distributions with zero mean,

$$\alpha_j|\delta_j,\psi_j \sim \mathcal{N}\left(0, r(\delta_j)\psi_j\right)\,, \qquad \psi_j|\boldsymbol{\vartheta} \sim p(\psi_j|\boldsymbol{\vartheta})\,,$$

where

$$r(\delta_j) = \begin{cases} r & \text{if } \delta_j = 0 \\ 1 & \text{if } \delta_j = 1 \end{cases}$$

and the distribution of ψ_j may depend on a further parameter $\boldsymbol{\vartheta}$. In particular, we consider normal spikes and slabs with constant $\psi_j \equiv V$ (called SSVS prior) and normal mixtures of inverse Gamma distributions (NMIG prior), where $\psi_j \sim \mathcal{G}^{-1}(\nu, Q)$. Priors with normal spikes and slabs were introduced in George and McCulloch (1993) to perform stochastic search variable selection and NMIG spikes and slabs were proposed in Ishwaran and Rao (2003) and Ishwaran and Rao (2005) for variable selection in Gaussian regression models and used in Konrath et al. (2008) for survival data. Note that for the NMIG prior marginally both spike and slab component are student distributions,

$$p_{\text{spike}}(\alpha_j) = t_{2\nu}(0, rQ/\nu) \qquad \text{and} \qquad p_{\text{slab}}(\alpha_j) = t_{2\nu}(0, Q/\nu)\,.$$

2.2.2 Dirac Spike

A Dirac spike is specified as $p_{\text{spike}}(\alpha_j) = p(\alpha_j|\delta_j = 0) = \Delta_0(\alpha_j)$. We combine Dirac spikes with slab components of the form

$$p_{\text{slab}}(\boldsymbol{\alpha}_{\boldsymbol{\delta}}) = f_N(\boldsymbol{\alpha}_{\boldsymbol{\delta}}; \mathbf{a}_{0,\boldsymbol{\delta}}, \mathbf{A}_{0,\boldsymbol{\delta}}\sigma^2)\,,$$

where $f_N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the multivariate $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-distribution. In particular we use

- the independence slab (i-slab), where $\mathbf{a}_{0,\boldsymbol{\delta}} = \mathbf{0}$ and $\mathbf{A}_{0,\boldsymbol{\delta}} = c\mathbf{I}$,
- the g-slab, where $\mathbf{a}_{0,\boldsymbol{\delta}} = \mathbf{0}$ and $\mathbf{A}_{0,\boldsymbol{\delta}} = g(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{-1}$,
- the fractional slab (f-slab), where $\mathbf{a}_{0,\boldsymbol{\delta}} = (\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{-1}\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{y}_c$ and $\mathbf{A}_{0,\boldsymbol{\delta}} = 1/b\,(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{-1}$.

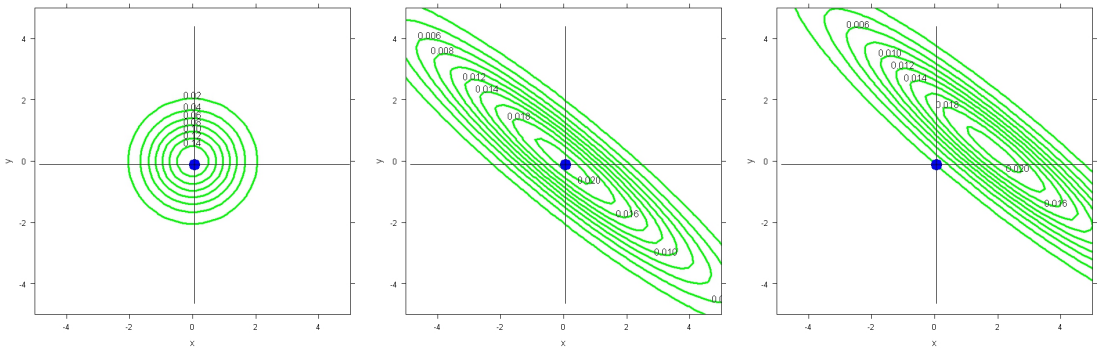

Figure 1: Contour plot of different priors for 2 regressors for $\boldsymbol{\delta} = (1, 1)$ and $\boldsymbol{\delta} = (0, 0)$: Dirac/i-slab (left), Dirac/g-slab (middle), Dirac/f-slab (right)

$\mathbf{X}_{\boldsymbol{\delta}}$ is the design matrix consisting only of those columns of $\mathbf{X}$ corresponding to non-zero effects, i.e. where $\delta_j = 1$. The g-slab is Zellner's g-prior (Zellner, 1986) for these effects and the f-slab is the corresponding fractional prior (O'Hagan, 1995). The idea of the fractional prior is to use a fraction b of the likelihood to determine a prior distribution for the parameters. In our specification the f-slab is not a fraction of the whole likelihood, but only of the part containing information on the regression coefficients $\boldsymbol{\alpha}$. Note that in contrast to the i-slab, regression coefficients α_j are not independent conditional on $\boldsymbol{\delta}$ for g- and f-slab, where the joint distribution of all effects with $\delta_j = 1$ is specified with a variance-covariance matrix equal to a scalar multiple of the Fisher information matrix. However, their mean is different: the g-slab is centered at the null vector, whereas the mean of f-slab is the LS estimate of the regression effects with $\delta_j = 1$. Figure 1 illustrates the differences between the three priors for two regressors showing the contours for the slab component for $\boldsymbol{\delta} = (1, 1)$ together with the position of the spike for $\boldsymbol{\delta} = (0, 0)$.

3 Inference

For both types of spike and slab priors posterior inference is feasible using MCMC methods, where the model parameters $(\mu, \boldsymbol{\delta}, \boldsymbol{\alpha}, \omega, \sigma^2)$ and additionally, under the NMIG prior, the scale parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)$ are sampled from their conditional posteriors. Depending on the type of the spike component, different sampling schemes have to be used: Whereas for an absolutely continuous spike the indicators δ_j can be sampled conditionally on the effects α_j, for a Dirac spike it is essential to draw $\boldsymbol{\delta}$ from the marginal posterior $p(\boldsymbol{\delta}|\mathbf{y})$ integrating over the parameters subject to selection, see Geweke (1996) and Smith and Kohn (1996). This requires evaluation of marginal likelihoods in each MCMC iteration. In normal regression models with conjugate priors (which are used here) analytical integration over the regression effects is feasible and hence marginal likelihoods can be computed rather cheaply. Details of the MCMC sampling schemes are given in the following two subsections.

3.1 MCMC for Absolutely Continuous Spikes

For priors with an absolutely continuous spike the full conditional distribution of $(\boldsymbol{\delta}, \boldsymbol{\psi})$ is given as

$$p(\boldsymbol{\delta}, \boldsymbol{\psi}|\boldsymbol{\alpha}, \omega, \mu, \sigma^2, \mathbf{y}) \propto \prod_{j=1}^{d} p(\alpha_j|\delta_j, \psi_j)p(\delta_j|\omega)p(\psi_j)p(\omega) \propto \prod_{j=1}^{d} p(\psi_j|\delta_j, \alpha_j)p(\delta_j|\alpha_j, \omega)\,.$$

Therefore, $\boldsymbol{\delta}$ and $\boldsymbol{\psi}$ can be sampled together in one block and the sampling scheme involves the following steps:

(1.) Sample μ from its posterior $\mu|\sigma^2, \mathbf{y} \sim \mathcal{N}(\bar{y}, \sigma^2/N)$.

(2.) Sample $\boldsymbol{\delta}$ and $\boldsymbol{\psi}$:

(2a.) For $j = 1, \ldots, d$ sample δ_j from

$$p(\delta_j = 1|\alpha_j, \omega) = \frac{1}{1 + \dfrac{1-\omega}{\omega} L_j}\,, \qquad L_j = \frac{p_{\text{spike}}(\alpha_j)}{p_{\text{slab}}(\alpha_j)}\,.$$

(2b.) For normal spikes and slabs, set $\psi_j \equiv V$. For student spikes and slabs, where $\psi_j \sim \mathcal{G}^{-1}(\nu, Q)$, sample ψ_j from its conditional posterior

$$\psi_j|\delta_j, \alpha_j \sim \mathcal{G}^{-1}\left(\nu + \frac{1}{2}, Q + \frac{\alpha_j^2}{2r(\delta_j)}\right)\,.$$

(3.) Sample ω from $\omega \sim \mathcal{B}(a_\omega + d_1, b_\omega + d - d_1)$ where $d_1 = \sum \delta_j$.

(4.) Sample $\boldsymbol{\alpha}$ from the normal posterior $\mathcal{N}(\mathbf{a}_N, \mathbf{A}_N)$ where $\mathbf{A}_N^{-1} = \frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{D}^{-1}$ and $\mathbf{a}_N = \mathbf{A}_N\mathbf{X}'\mathbf{y}_c$. $\mathbf{D}$ is a diagonal matrix with entries $r(\delta_j)\psi_j$, $j = 1, \ldots, d$.

(5.) Sample the error variance σ^2 from the posterior $\sigma^2|\mathbf{y}_c, \boldsymbol{\alpha} \sim \mathcal{G}^{-1}(s_N, S_N)$, where $s_N = (N-1)/2$ and $S_N = \frac{1}{2}(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})'(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})$.

3.2 Sampling Steps for a Dirac Spike

For a Dirac spike, $\delta_j = 0$ implies $\alpha_j = 0$ and vice versa. To avoid reducibility of the Markov chain, it is essential to draw $\boldsymbol{\delta}$ from the marginal posterior

$$p(\boldsymbol{\delta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\delta})p(\boldsymbol{\delta})\,,$$

where effects subject to selection are integrated out. Here $p(\mathbf{y}|\boldsymbol{\delta})$ denotes the marginal likelihood of the linear regression model (1) with design matrix $\mathbf{X}_{\boldsymbol{\delta}}$. For Dirac spikes combined with i-, g- or f-slab on $\boldsymbol{\alpha}_{\boldsymbol{\delta}}$ the marginal likelihood can be derived analytically as

$$p(\mathbf{y}|\boldsymbol{\delta}) = \frac{1}{\sqrt{N}(2\pi)^{(N-1)/2}} \frac{|\mathbf{A}_{\boldsymbol{\delta}}|^{1/2}}{|\mathbf{A}_{0,\boldsymbol{\delta}}|^{1/2}} \frac{\Gamma(s_N)}{S_N^{s_N}}\,, \tag{4}$$

where $s_N = (N-1)/2$ and $S_N = \frac{1}{2}(\mathbf{y}_c'\mathbf{y}_c - \mathbf{a}_{\boldsymbol{\delta}}'\mathbf{A}_{\boldsymbol{\delta}}^{-1}\mathbf{a}_{\boldsymbol{\delta}})$. $\mathbf{a}_{\boldsymbol{\delta}}$ and $\mathbf{A}_{\boldsymbol{\delta}}$ are parameters of the posterior of $\boldsymbol{\alpha}_{\boldsymbol{\delta}}$: $\mathbf{A}_{\boldsymbol{\delta}} = ((\mathbf{X}_{\boldsymbol{\delta}}'\mathbf{X}_{\boldsymbol{\delta}}) + \frac{1}{c}\mathbf{I})^{-1}$ for the i-slab, $\mathbf{A}_{\boldsymbol{\delta}} = \frac{g}{g+1}(\mathbf{X}_{\boldsymbol{\delta}}'\mathbf{X}_{\boldsymbol{\delta}})^{-1}$ for the

g-slab and $\mathbf{A}_{\boldsymbol{\delta}} = (\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{-1}$ for the f-slab; the posterior mean is $\mathbf{a}_{\boldsymbol{\delta}} = \mathbf{A}_{\boldsymbol{\delta}}\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{y}_c$ for any of the three slabs. Details are given in Appendix A.

With this marginalization it is possible to sample the parameters $\boldsymbol{\delta}$, σ^2 and μ in one block. Hence, the MCMC scheme for Dirac spikes involves the following steps:

(1.) Sample $(\boldsymbol{\delta}, \sigma^2, \mu)$ from the posterior $p(\boldsymbol{\delta}|\mathbf{y})p(\sigma^2|\mathbf{y}, \boldsymbol{\delta})p(\mu|\mathbf{y}, \boldsymbol{\delta}, \sigma^2)$.

(1a.) Sample each element δ_j of the indicator vector $\boldsymbol{\delta}$ separately from $p(\delta_j = 1|\boldsymbol{\delta}_{\setminus j}, \mathbf{y})$ given as

$$p(\delta_j = 1|\boldsymbol{\delta}_{\setminus j}, \mathbf{y}) = \frac{1}{1 + \dfrac{1-\omega}{\omega} R_j}, \qquad R_j = \frac{p(\mathbf{y}|\delta_j = 0, \boldsymbol{\delta}_{\setminus j})}{p(\mathbf{y}|\delta_j = 1, \boldsymbol{\delta}_{\setminus j})}.$$

Here $\boldsymbol{\delta}_{\setminus j}$ denotes the vector $\boldsymbol{\delta}$ consisting of all elements of $\boldsymbol{\delta}$ except δ_j. Elements of $\boldsymbol{\delta}$ are updated in a random permutation order.

(1b.) Sample the error variance σ^2 from the $\mathcal{G}^{-1}(s_N, S_N)$-distribution.

(1c.) Sample the mean μ from the $\mathcal{N}(\bar{y}, \sigma^2/N)$-distribution.

(2.) Sample ω from $\omega \sim \mathcal{B}(a_\omega + d_1, b_\omega + d - d_1)$, where $d_1 = \sum \delta_j$.

(3.) Set $\alpha_j = 0$ if $\delta_j = 0$. Sample the non-zero elements $\boldsymbol{\alpha}_{\boldsymbol{\delta}}$ from the normal posterior $\mathcal{N}(\mathbf{a}_{\boldsymbol{\delta}}, \mathbf{A}_{\boldsymbol{\delta}}\sigma^2)$.

For both g- and f-slab, the posterior variance covariance matrix $\mathbf{A}_{\boldsymbol{\delta}}$ is a scalar multiple of the prior variance covariance matrix $\mathbf{A}_{0,\boldsymbol{\delta}}$. Thus for computing the marginal likelihood (4), the determinant of $\mathbf{A}_{\boldsymbol{\delta}}$ is not required which speeds up sampling compared to i-slabs.

4 Simulated Data

We investigate performance of the different MCMC implementations for simulated data. Interest lies in correct selection of regressors as well as sampling efficiency of posterior inclusion probabilities. We expect draws of the posterior probabilities $p^{(m)}(\delta_j = 1)$, $m = 1, \ldots, M$ to have higher autocorrelations for continuous than for Dirac spikes. It is however not obvious which implementation will have higher computational cost in CPU time: With a Dirac spike only coefficients with $\delta_j = 1$ have to be sampled, as those with $\delta_j = 0$ are restricted exactly to zero, whereas for a continuous spike the dimension of the model is not reduced during MCMC. On the other hand, specifying a continuous spike will save CPU time as no marginal likelihoods have to be computed.

To investigate correct model selection we simulate 100 data sets with $N = 40$ observations from a linear regression model with mean $\mu = 1$ and $\sigma^2 = 1$ and nine covariates. We consider two setups for the covariate vectors $\mathbf{x}_j$, which are drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{C})$-distribution: independent regressors, where $\mathbf{C} = \mathbf{I}$, and correlated regressors generated as in Tibshirani (1996), where $\mathbf{C}$ is a correlation matrix with $C_{jk} = \rho^{|j-k|}$ with $\rho = 0.8$. For both independent and correlated regressors we set three regression effects to each of the values "2" (strong effects), "0.2" (weak effects) and "0" (zero effects).

In the simulation studies, we use an uninformative $\mathcal{B}(1, 1)$-prior for ω. To mimic Dirac spikes closely, a small variance ratio r of continuous spikes and slabs would be preferred, however r should not be too small to avoid MCMC getting stuck in the spike

component. Following the recommendations in George and McCulloch (1993) we set $r = 1/10000$.

It is well known that the choice of the slab distribution is critical for model selection. Our choice for the slab variance is motivated by the fact that model selection based on Bayes factors is consistent for the g-prior with $g = N$ (see Fernández, Ley, and Steel, 2001). Hence we choose $g = N$ and match the variances of the other slabs to equal the variance of the g-slab if regressors are orthogonal, i.e. we choose $b = 1/N$ and $V = 1$. For the NMIG-prior we choose $\nu = 5$, which corresponds to a t-distribution with 10 degrees of freedom, and $Q = 4$.

For each data set, MCMC was run for $M = 5000$ iterations after a burn-in of 1000 draws. The first 500 draws of the burn-in were drawn from the full model including all regressors.

4.1 Model Selection Performance

Posterior inclusion probabilities are estimated by their posterior mean, i.e. the average of the inclusion probabilities $p^{(m)}(\delta_j = 1)$ in the MCMC iterations. Figure 2 shows box-plots of these estimates in the 100 simulated data sets with independent regressors. Regressors with strong effect are perfectly classified with estimated posterior inclusion probabilities being equal to 1 (rounded) in all 100 data sets. Variation of the estimated posterior inclusion probabilities is high for regressors with weak and zero effect which indicates that regression coefficients of smaller magnitude are hard to classify. Posterior inclusion probabilities tend to be slightly smaller for the Dirac/i-slab and the Dirac/g-slab priors than for the other priors.

For orthogonal regressors Barbieri and Berger (2004) showed that the median probability model, i.e. the model including regressors with posterior inclusion probability larger than 0.5, is the best model with regard to predictive performance. Table 1 reports the number of data sets where each of the regressors with weak or zero effect is included in the median probability model. Results are not shown for regressors with strong effect as these are included in all 100 data sets under any prior. Whereas under the Dirac spike combined with i- or g-slab classification is better for zero effects, weak effects are detected less often than under the other three priors. Overall performance is similar for all priors with mean misclassification rates (computed over weak and zeros effects) from 41.6 % (Dirac/f-slab) to 43.3 % (NMIG).

Figure 3 shows the estimated posterior inclusion probabilities for simulated data with correlated regressors. The order of regressors with strong, weak and zero effects is different now, to get insight in the effects of correlations which are highest for neighboring regressors. Posterior inclusion probabilities of regressors with strong effects show more variation than for independent regressors but are close to 1 in almost all cases. A pronounced difference however occurs for regressors with weak and zero effects, which are slightly smaller for the Dirac/g-slab and Dirac/f-slab prior but considerably higher for priors with independent slabs (Dirac/i-slab, SSVS and NMIG) than in Figure 2. As a consequence, regressors with weak and zero effects are included in the median probability model less often under the Dirac/g-slab and Dirac/f-slab prior but more often under priors with independent slabs, when regressors are correlated. Table 2 reports in how many data

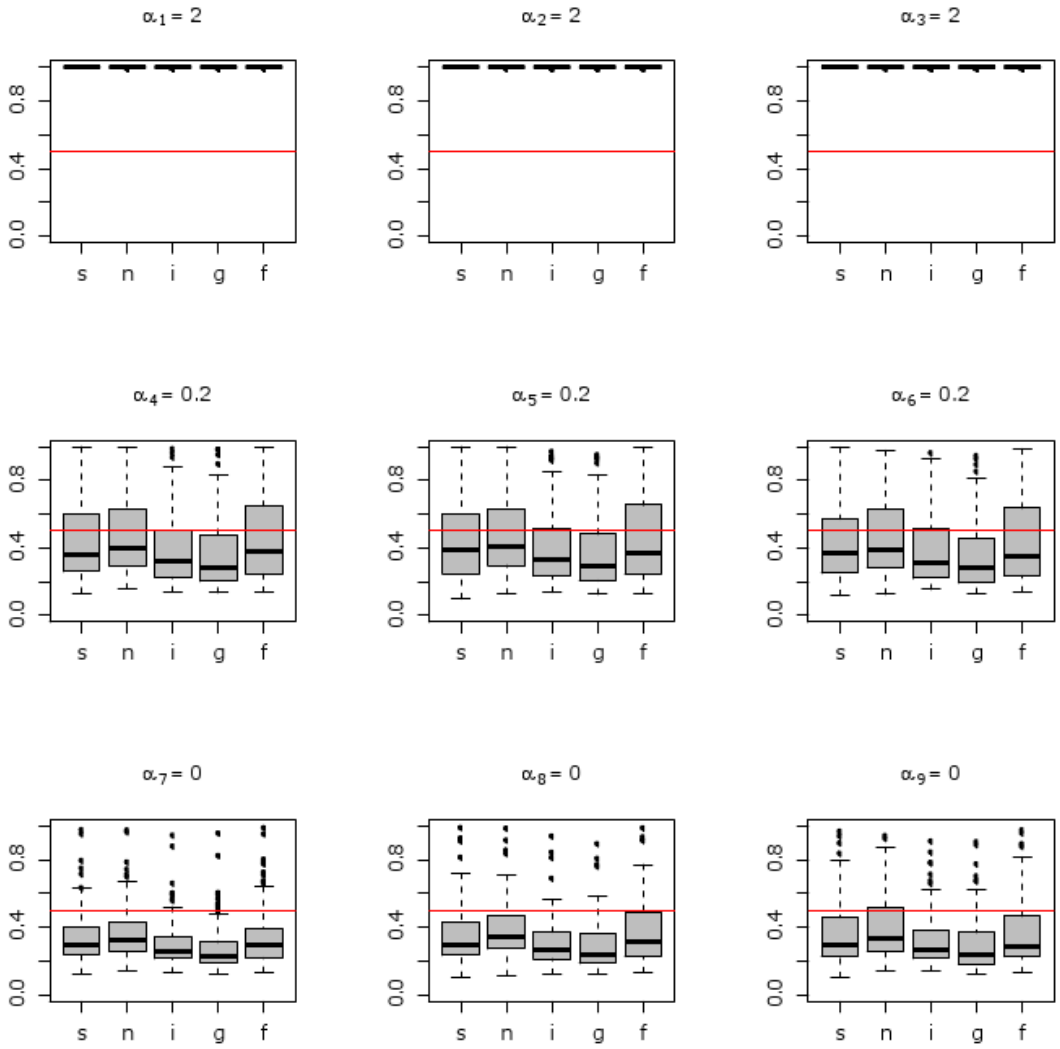

Figure 2: Independent regressors: Posterior inclusion probabilities of each regressor for 100 simulated data sets (s=SSVS prior, n=NMIG prior, i=Dirac/i-slab, g=Dirac/g-slab, f=Dirac/f-slab)

sets each regressor is included in the median probability model. As estimated posterior inclusion probabilities of regressors with strong effects are higher than 0.5 in all data sets (except in one data set for the g-slab), only results for weak and zero effects are given. Mean misclassification rates (computed over weak and zeros effects) are higher than for independent regressors, but again very similar, ranging from 47.5 % (Dirac/f-slab) to 48 % (Dirac/i-slab). Obviously the correlation structure of the prior has an effect on the posterior inclusion probability when regressors are correlated. We will return to this issue in Section 5.2, where we investigate this effect analytically, though in a simpler setting.

Further simulations carried out in Malsiner-Walli (2010) indicate that posterior inclusion probabilities depend on the variance of the slab component: Posterior inclusion

Table 1: Independent regressors: Number of data sets where $\hat{p}(\delta_j = 1) > 0.5$.

		Continuous spike		Dirac spike		
j	α_j	SSVS	NMIG	i-slab	g-slab	f-slab
4	0.2	31	36	25	23	36
5	0.2	33	35	26	25	37
6	0.2	28	32	26	23	38
7	0	12	15	11	9	15
8	0	18	22	11	8	24
9	0	21	26	13	11	22

Table 2: Correlated regressors: Number of data sets where $\hat{p}(\delta_j = 1) > 0.5$

		Continuous spike		Dirac spike		
j	α_j	SSVS	NMIG	i-slab	g-slab	f-slab
3	0	62	66	58	6	19
5	0.2	66	73	60	6	22
6	0	60	66	44	5	26
7	0	55	63	48	2	18
8	0.2	67	73	50	10	26
9	0.2	57	63	52	10	30

probabilities decrease with increasing slab variance. This is another issue which we investigate analytically in Section 5 and illustrate in the application in Section 6.

4.2 Comparing Sampling Efficiencies

As MCMC draws are correlated, it is of interest to compare MCMC implementations for the different priors with respect to their sampling efficiency. Table 3 reports mean inefficiency factors (also called integrated autocorrelation times) for regressors with weak and zero effects. Inefficiency factors, defined as $\tau = 1 + 2\sum_{l=1}^{L} \rho(l)$, where $\rho(l)$ is the empirical autocorrelation at lag l, were computed using the initial monotone sequence estimator (Geyer, 1992) for L. If inclusion probabilities $p^{(m)}(\delta_j = 1)$ are numerically equal to 1 in all iterations, inefficiency factors cannot be computed. This occurred for one effect in one data set under the SSVS prior and hence the average reported in Table 3 for the SSVS prior is based only on the remaining posterior inclusion probabilities. Interestingly for correlated regressors inefficiency factors are lower for the Dirac/g-slab and Dirac/f-slab prior and higher for priors with independent slab. This might result from the decrease/increase of posterior inclusion probabilities: Wagner and Duller (2011) also observed smaller inefficiency factor for low inclusion probabilities, though in logit models.

As expected, inefficiency factors are considerably higher for priors with continuous spikes than for Dirac spikes. Further simulations in Malsiner-Walli (2010) showed that, for continuous spikes, the choice of the variance ratio r as well as the actual implemen-

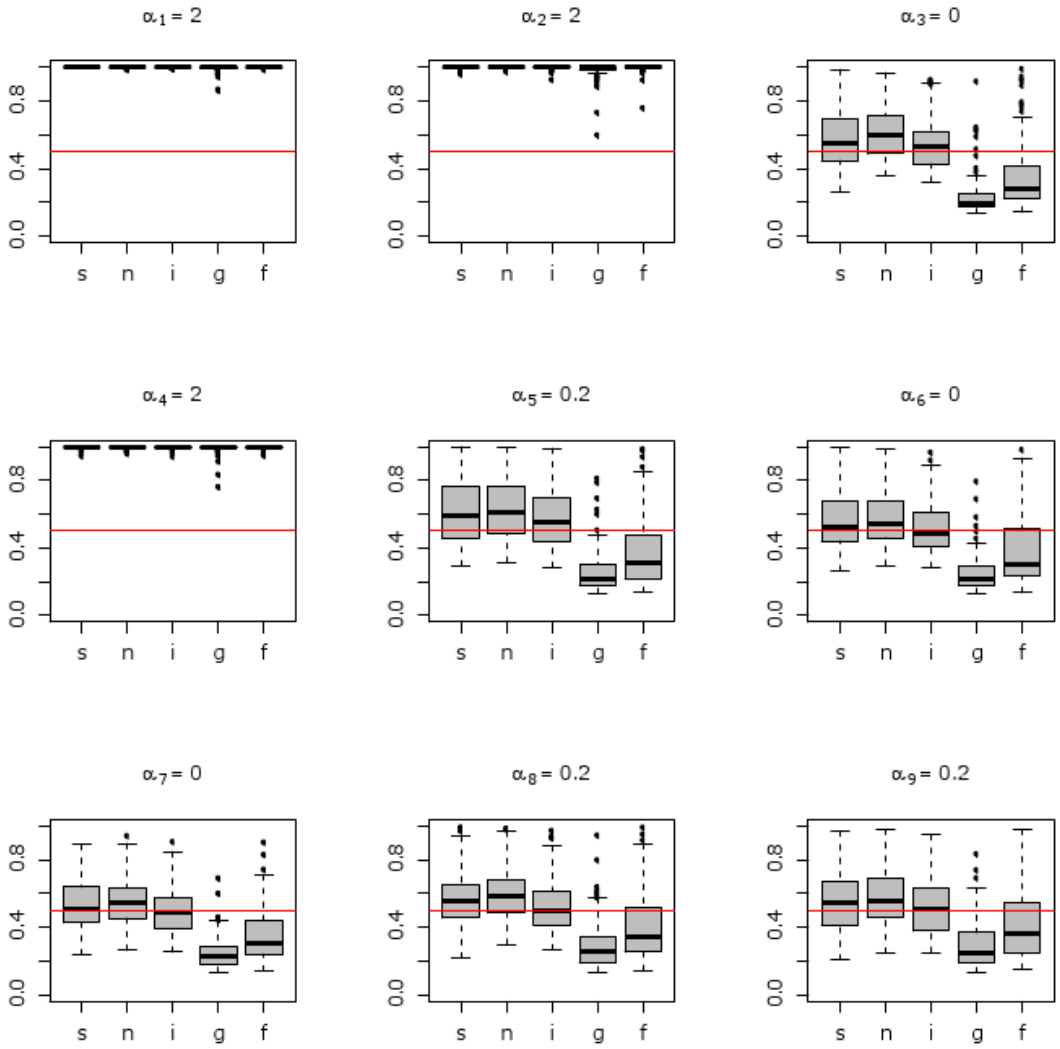

Figure 3: Correlated regressors: Posterior inclusion probabilities of each regressor for 100 simulated data sets (s=SSVS prior, n=NMIG prior, i=Dirac/i-slab, g=Dirac/g-slab, f=Dirac/f-slab).

tation can have an impact on sampling efficiency: Autocorrelations and inefficiency are lower for higher values of r, e.g. $r = 1/1000$ yields similar estimates for posterior inclusion probabilities but with less autocorrelated draws. Under the NMIG prior, posterior inclusion probabilities could be computed alternatively conditional on the variance parameters ψ_j as in Konrath et al. (2008), which however leads to considerably higher autocorrelations than using the marginal t-distribution.

To assess sampling efficiency with computing time taken into account, Table 4 reports effective sample sizes per second averaged over weak and zero effects. The effective sample size $ESS = M/\tau$ estimates the number of independent samples required to obtain a parameter estimate with the same precision as the MCMC estimate. Results in Table 4

Table 3: Averaged inefficiency factors

	Continuous spike		Dirac spike		
Regressors	SSVS	NMIG	i-slab	g-slab	f-slab
Independent	26.3	23.7	3.3	3.1	3.2
Correlated	30.1	27.2	3.7	2.5	2.9

Table 4: Averaged effective sample size per sec.

	Continuous spike		Dirac spike		
Regressors	SSVS	NMIG	i-slab	g-slab	f-slab
Independent	33.3	27.6	16.6	34.3	23.2
Correlated	25.1	18.9	14.9	43.3	27.1

are based on all MCMC chains, where inefficiency factors could be computed. Though sampling efficiency is much higher for priors with Dirac spikes differences in effective sample sizes are much less pronounced and priors with absolutely continuous spikes perform roughly similar to Dirac/g- and Dirac/f-slab. Even in this rather low-dimensional model with only nine regressors, computational cost for the Dirac/i-slab prior is too high to be outweighed by the smaller inefficiency factors. Due to lower inefficiency factors priors with g- and f-slabs have even higher ESS/sec for correlated regressors.

5 Posterior Inclusion Probabilities

Results of the simulation study indicate that posterior inclusion probabilities largely depend on the slab distribution. To get further insight into the effect of different slabs we investigate the inclusion probability of one regressor $\mathbf{x}_j$ conditional on $\boldsymbol{\delta}_{\setminus j}$ for priors with Dirac spikes (i.e. the Dirac/i-slab, Dirac/g-slab and Dirac/f-slab prior) in two simple special cases: for orthogonal regressors and in a model with only two correlated regressors. For simplicity we assume that the error variance σ^2 is known. Details on the computation of posterior inclusion probabilities are given in Appendix B. We will denote by $s_y^2 = \frac{1}{N}\mathbf{y}_c'\mathbf{y}_c$ the sample variance of $\mathbf{y}$, by r_{yj} the sample correlation between $\mathbf{y}$ and $\mathbf{x}_j$ and by $s_j^2 = \frac{1}{N}\mathbf{x}_j'\mathbf{x}_j$ the sample variance of covariate $\mathbf{x}_j$.

5.1 Orthogonal Regressors

For orthogonal regressors, i.e. $\mathbf{X}'\mathbf{X} = \mathrm{diag}(Ns_j^2)$, $j = 1, \ldots, d$, the posterior inclusion probability of $\mathbf{x}_j$ can be written as a function of the LS-estimate $\hat{\alpha}_j = \frac{r_{yj}s_y}{s_j}$ as

$$p(\delta_j = 1|\mathbf{y}, \boldsymbol{\delta}_{\setminus j}) = \frac{1}{1 + \exp(h(\hat{\alpha}_j, \theta)/2)\dfrac{(1-\omega)}{\omega}}, \tag{5}$$

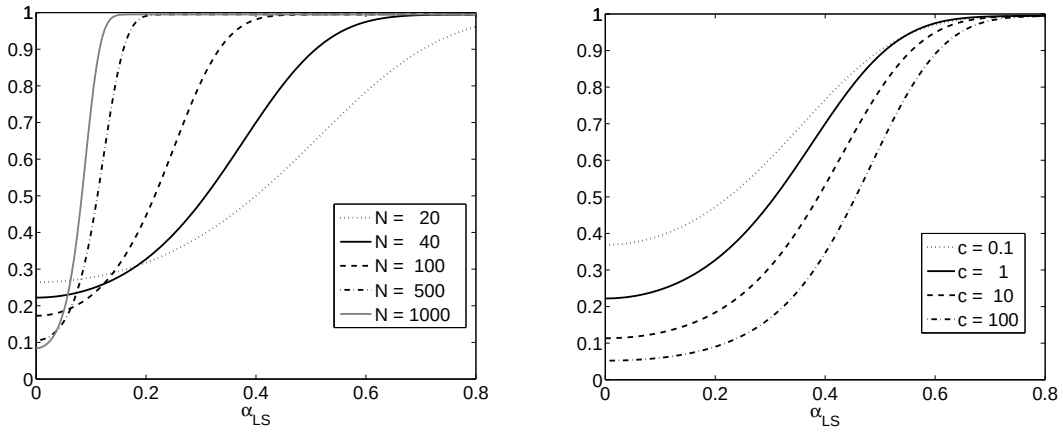

Figure 4: Independent regressors: Posterior inclusion probability under the Dirac/i-slab prior (for $\sigma^2 = 1$, integrated over ω). Left: $c = 1$, different values of N; right: $N = 40$, different values of c.

where θ is the variance parameter of the slab distribution, i.e. c for the i-slab, g for the g-slab and b for the f-slab. Under any of the three slabs, the inclusion probability does not depend on $\boldsymbol{\delta}_{\setminus j}$. In particular we obtain (see Appendix B.1)

$$\text{i-slab:} \quad h(\hat{\alpha}_j, c) = -N\frac{\hat{\alpha}_j^2 s_j^2}{\sigma^2}\frac{1}{1+1/(Ns_j^2 c)} + \log(Ns_j^2 c + 1)\,, \tag{6}$$

$$\text{g-slab:} \quad h(\hat{\alpha}_j, g) = -\frac{N\hat{\alpha}_j^2 s_j^2}{\sigma^2}\frac{g}{g+1} + \log(g+1)\,, \tag{7}$$

$$\text{f-slab:} \quad h(\hat{\alpha}_j, b) = -\frac{N\hat{\alpha}_j^2 s_j^2}{\sigma^2}(1-b) - \log(b)\,. \tag{8}$$

In formulas (6) – (8) the first term is proportional to $N\frac{\hat{\alpha}_j^2}{\sigma^2}$ which, following Dey, Ishwaran, and Rao (2008), can be interpreted as the signal of the regression coefficient contained in the data. Hence, posterior inclusion probabilities increase with both, sample size N and the size of the estimated effect $|\hat{\alpha}_j^2|$. The second term can be interpreted as a penalty term: It increases with the slab variance, and hence the posterior inclusion probability decreases as a function of the slab variance. Figure 4 shows posterior inclusion probabilities under the i-slab as a function of the LS estimate $\hat{\alpha}$ for various samples sizes N and variances c. In contrast to i-slabs the penalty term does not depend on the scale of the regressor under g- and f-slabs. For standardized orthogonal regressors ($s_j^2 = 1$) posterior inclusion probabilities are identical for g-slab and i-slab when $Nc = g$ and slightly higher under the f-slab when $b = 1/g$. This corresponds to the simulation results, see Figure 2.

To illustrate the dependence of posterior inclusion probabilities on the effect signal $\hat{\alpha}$ we generated 100 data sets of size $N = 200$, with 21 regressors generated as independent standard normal random variables and effects from 0 to 0.4 in increments of 0.02. Posterior inclusion probabilities were estimated under the less restrictive assumption of unknown error variance using the MCMC scheme described in Section 3.2. Figure 5 shows estimated posterior inclusion probabilities for the Dirac/i-slab plotted versus $\hat{\alpha}$

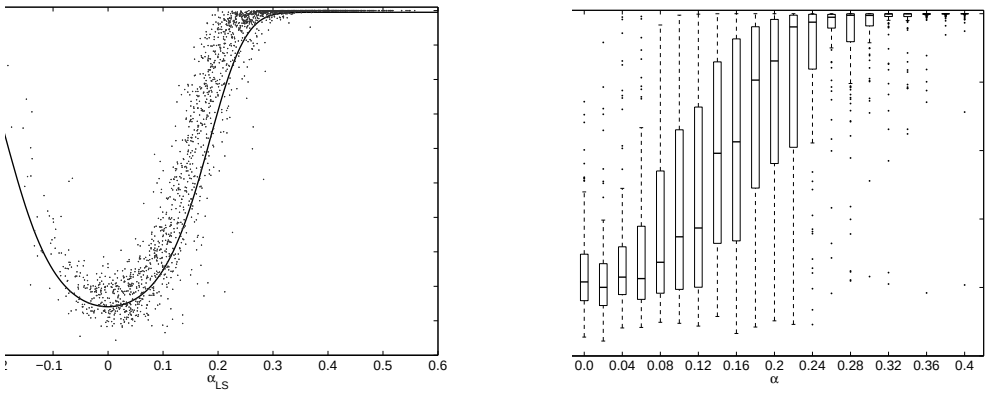

Figure 5: Simulated data: Posterior inclusion probabilities under Dirac/i-slab prior ($c = 1$, $N = 200$) as a function of the LS-estimate $\hat{\alpha}$ (left) and of the true effect α (right).

(left panel). Posterior inclusion probabilities do not exactly equal the theoretical values computed from formula (5), which are shown as a line. This is not surprising as the assumptions for the derivation of the formula are not met exactly: Firstly, due to stochastic variation regressors are not perfectly orthonormal and secondly, in the MCMC scheme the marginal likelihood is computed using formula (4) with marginalization over the error variance σ^2. In the right panel of Figure 5 estimated posterior inclusion probabilities are plotted against the "true" effect sizes α used for data generation. Conditional on α variation of the posterior inclusion probabilities is much higher as additionally the variation LS estimate $\hat{\alpha}$ is reflected.

5.2 Two Correlated Regressors

To investigate the effect of correlation between regressors we consider a model with only two standardized regressors $\mathbf{x}_1$ and $\mathbf{x}_2$ (i.e. $s_j^2 = 1$) and assume that $\mathbf{x}_1$ is included in the model, i.e. $\delta_1 = 1$. We denote by $r_{12} = \frac{1}{N}\mathbf{x}_1'\mathbf{x}_2$ the sample correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$ and by $\hat{\alpha}_2 = s_y(r_{y2} - r_{12}r_{y1})/(1 - r_{12}^2)$ the LS estimate of α_2 in the model including both regressors.

We are interested in the conditional posterior inclusion probability of $\mathbf{x}_2$, which can be written as a function of

$$h(\hat{\alpha}_2, \sim) = 2\big(\log p(\mathbf{y}|\delta_1 = 1, \delta_2 = 0, \sigma^2) - \log p(\mathbf{y}|\delta_1 = 1, \delta_2 = 1, \sigma^2)\big)$$

as in equation (5). Under g- and f-slab it is straightforward to derive $h(\hat{\alpha}_2, \sim)$ as

$$h(\hat{\alpha}_2, g) = -\frac{N\hat{\alpha}_2^2}{\sigma^2}(1 - r_{12}^2)\,\frac{g}{g+1} + \log(g+1)\,,$$
$$h(\hat{\alpha}_2, b) = -\frac{N\hat{\alpha}_2^2}{\sigma^2}(1 - r_{12}^2)(1-b) - \log(b)\,,$$

see Appendix B.2 for details. For a given value of the LS estimate $\hat{\alpha}_2$, the probability of including $\mathbf{x}_2$ additionally to $\mathbf{x}_1$ in the model therefore decreases with the square of the

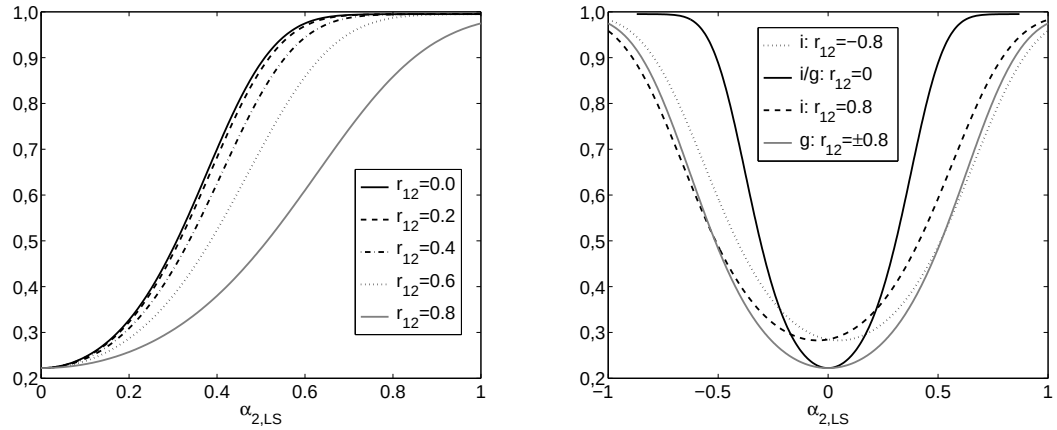

Figure 6: Correlated regressors: Posterior inclusion probability of regressor $\mathbf{x}_2$, conditional on $\delta_1 = 1$ (integrated over ω, $N = 40$). Left: g-slab, different values of r_{12}, right: comparing g- and i-slab for different values of r_{12} ($s_y = 2$, $r_{1y} = 0.9$).

correlation r_{12} between the two regressors. Figure 6 (left panel) shows the conditional posterior inclusion probabilities of $\mathbf{x}_2$ under the Dirac/g-slab as a function of $\hat{\alpha}_2$ for different values of r_{12}. Obviously, for highly correlated regressors the inclusion probability of the second regressor can be reduced dramatically.

For the Dirac/i-slab prior, simple but tedious algebra yields

$$h(\hat{\alpha}_2, c) = -\frac{N}{Q\sigma^2}\Big(\hat{\alpha}_2(1 - r_{12}^2) + \frac{r_{y2}s_y}{Nc}\Big)^2 + \log\Big(Nc(1 - r_{12}^2) + 1 + \frac{r_{12}^2}{1 + 1/(Nc)}\Big)\,,$$

where

$$Q = (1 - r_{12}^2) + \frac{1}{Nc}(3 - r_{12}^2) + \frac{3}{(Nc)^2} + \frac{1}{(Nc)^3}\,.$$

The first summand in the function $h(\hat{\alpha}_2, c)$ is different from the corresponding term for g- and f-slab. However, as it is dominated by $-\frac{N\hat{\alpha}_2^2}{\sigma^2}(1 - r_{12})$, this difference will vanish for increasing sample size N. Further, in contrast to g- and f-slab, the penalty term $\log(\sim)$ depends on the regressor correlation r_{12} leading to less penalization of the additional regressor $\mathbf{x}_2$ compared both to orthogonal regressors and to g- and f-slabs. Therefore, posterior inclusion probabilities under i-slabs will be higher for correlated regressors. The conditional inclusion probabilities of $\mathbf{x}_2$ under the i-slab depend not only on $\hat{\alpha}_2$, but also on r_{2y} and are no longer symmetric in $\hat{\alpha}_2$, at least for small sample size N. This is shown in Figure 6 (right panel), which compares the inclusion probability of $\mathbf{x}_2$ for g- and i-slab for different correlations r_{12}. Posterior inclusion probabilities are considerably smaller under the g-slab for small absolute values of $\hat{\alpha}_2$. Results from our simulations, see Figure 3, suggest that differences in posterior inclusion probabilities under i- and g-slab can be even more pronounced in models with more regressors.

6 Application

We illustrate application of the different variable selection methods on a data set of psychiatric patients. Metabolic disorders and weight gain are common problems and side effects of psychiatric medication. To investigate how bodyweight and parameters of lipid and glucose metabolism are influenced by psychiatric inpatient treatment, a prospective study was performed at a department of the Wagner-Jauregg hospital in Upper Austria from October 2003 to March 2004. Several lipid and glucose parameters, namely total cholesterol (`chol`), high density lipoprotein cholesterol (`hdl`), low density lipoprotein cholesterol (`ldl`), triglycerides (`nf`) and fasting glucose (`nbz`) were measured at admission and at discharge of the department. Medication, if any, was assessed as prescribed at discharge and assigned to 16 drugs or types of drugs. Additionally, several patient-related variables were collected: age, sex, height, smoking, body mass index at admission and duration of the stay.

The focus of our analysis is to identify covariates influencing the `change in HDL`, and we used the lipid and glucose values at admission, the 16 different drug types and all patient variables as potential regressors. Excluding observations with missing values, leaves data on 231 patients with 27 regressors for the analysis. Pairwise correlations between covariates are smaller than 0.1 in most cases, only three correlations are higher than 0.4 (`sex` and `height`: $r = 0.67$; `chol_admiss` and `ldl_admiss`: $r = 0.86$ and `drug A` and `drug B`: $r = 0.89$).

Following Gelman, Jakulin, Pittau, and Su (2008), metric covariates were standardized, and dummy covariates were centered. As a first step an exploratory Bayesian analysis of the unrestricted model under the prior $\mathcal{N}(\mathbf{0}, c\mathbf{I})$ with $c = 5$ was carried out. Figure 7 shows the posterior estimates and 95 %-credible intervals of the regression effects. Only for 6 covariates (covariates number 6: `chol_admiss`, 8: `hdl_admiss`, 9: `ldl_admiss`, 16: `drug F`, 20: `drug J` and 27: `bmi_admiss`) these intervals do not contain zero, indicating that the corresponding effects "significantly" differ from zero.

As a next step, MCMC for variable selection was run for $M = 5000$ iterations (after a burn-in of 1000, with the first 500 draws of the burn-in drawn from the unrestricted model) for Dirac spike priors and $M = 50000$ iterations (after 10000 burn-in with the first 5000 draws from the unrestricted model) for priors with absolutely continuous spikes. To match the slab variances the response was standardized with the estimated residual standard deviation ($s = 15.4$) of the full regression model. Hyper-parameters were chosen as in the simulation studies: we used a variance ratio of $r = 1/10000$, $\nu = 5$ and $c = 1$ and the other parameters were set to $g = Nc$, $b = 1/g$, $V = c$ and $Q = 4$.

Posterior inclusion probabilities were roughly equal for all covariates under the Dirac/i-slab, the SSVS and the NMIG prior, however, considerably smaller for Dirac/g- and Dirac/f-slab priors. Table 5 reports estimated posterior inclusion probabilities for the covariates selected in the median probability model under the Dirac/i-slab prior. Results correspond well with the exploratory analysis of the unrestricted model: the selected covariates build a subset of those identified as having a "significant" effect, and in contrast to the exploratory analysis, Bayesian variable selection automatically controls for multiple-testing.

From the medical point of view the goal of the analysis was to obtain a classification of

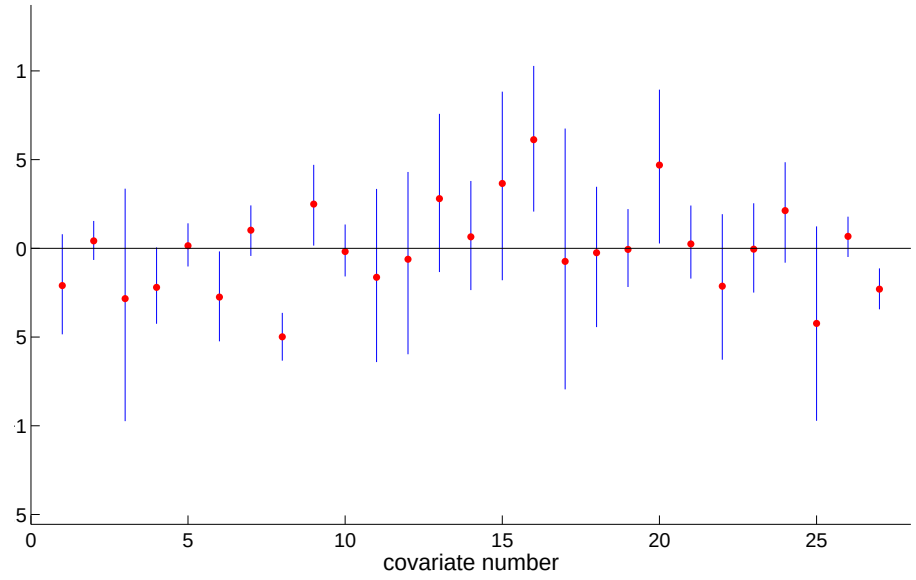

Figure 7: HDL data: Posterior means and 95%- credible intervals for regression effects in the unrestricted model

Table 5: HDL data: Posterior inclusion probabilities (for $c = 1$)

	Continuous spike		Dirac spike		
Covariate number	SSVS	NMIG	i-slab	g-slab	f-slab
8 (`hdl_admiss`)	1.00	1.00	1.00	1.00	1.00
16 (`drug F`)	0.78	0.82	0.81	0.49	0.49
20 (`drug J`)	0.63	0.61	0.68	0.29	0.29
27 (`bmi_admiss`)	0.56	0.53	0.62	0.34	0.32

covariates into those which have nearly zero effect and can be excluded from the model and others which eventually affect the response variable. Therefore, variable selection was not based on the Dirac/g- and f-slab-priors which more heavily penalize dependent regressors than independent slabs.

Table 6 shows inefficiency factors and effective sample size per sec. averaged over all covariates (except covariate 8). Again inefficiency factors of the posterior inclusion probabilities are considerably higher under priors with continuous spikes. However, when computational effort is taken into account again all priors except Dirac/i-slab prior perform similar.

Finally, to study the effect of the slab variance, we ran MCMC for different values $c = 1, 2.5, 5, 10$ for the i-slab and corresponding parameters of the other priors. The resulting posterior inclusion probability paths shown in Figure 8 for the Dirac/i-slab and Dirac/g-slab priors, demonstrate the effect of increasing penalization of regressors for larger slab variances.

Table 6: HDL data: Sampling efficiency of posterior inclusion probabilities

	Continuous spike		Dirac spike		
	SSVS	NMIG	i-slab	g-slab	f-slab
Averaged inefficiency factor	57.7	43.2	3.1	2.1	2.3
Averaged effective sample size/sec.	15.6	12.8	5.9	16.9	10.9

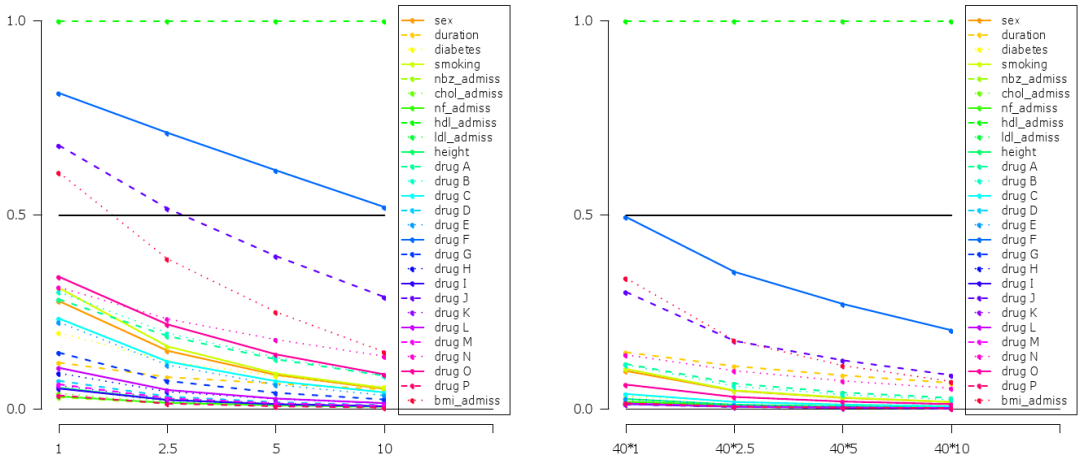

Figure 8: HDL data: Posterior inclusion probability paths for different slab variances c for the Dirac/i-slab prior (left) and different values of g for the Dirac/r-slab prior (right)

7 Summary and Discussion

We compared different spike and slab priors which are widely used for Bayesian variable selection. Simulation studies suggest that for orthogonal regressors different priors act rather similar when the slab variances are matched, which is confirmed by theoretical results for Dirac spike priors (and known error variance). The posterior inclusion probability of a specific regressor increases with the signal of the effect in the data and decreases with the variance of the slab component. Compared to orthogonal regressors, both simulations as well as theoretical results, indicate that for a given effect signal in the data, posterior inclusion probabilities are smaller under g-and f-slabs and higher for priors with independent slabs if regressors are correlated. This result suggests to use g- or f-slabs in practical applications where interest is in avoiding "false positives" and independent slabs either with Dirac or continuous spikes if the goal is not to miss eventually important predictors.

From a computational point of view, priors with continuous spikes are a fast alternative to the Dirac/i-slab prior as higher autocorrelations are outweighed by less computation time. Mixing of the sampler is better for the NMIG than the SSVS prior at the cost of a small additional computational effort. MCMC getting stuck at $p(\delta_j = 1) = 1$ is more severe for SSVS than NMIG priors, where it occurred only for regressors with strong

effects.

A drawback of all priors considered here is that they do not well discriminate between regressors with zero and weak effects. Choosing a smaller variance for the slab component does not solve this problem as inclusion probabilities of all effects, even of zero effects, will increase. For Bayesian testing, (Johnson and Rossell, 2010) recently proposed so called non-local prior densities, which are zero in the parameter space of the null hypothesis to facilitate separation between null and the alternative. Spike and slab priors compared in this paper could be modified in this direction with slab components having a mode different from zero. Prior information on the size of "relevant" effects could be incorporated by specifying either one slab or, if no information on the effect sign is available, two slabs with a positive and a negative mode, respectively. For slabs which are normal or NMIG, MCMC schemes presented in this work could be used with slight modifications.

Acknowledgements

The authors thank Univ. Doz. Prim. Dr. Hans Rittmannsberger (Wagner-Jauregg-Krankenhaus Linz) for providing the data and many helpful comments. We would also like to thank the anonymous referee for his suggestions to improve the paper and Christoph Pamminger for careful reading of the manuscript.

Appendix

A Marginal Likelihoods

We consider the normal regression model (1) with $N \times d$ regressor matrix $\mathbf{X}$ with centered columns, i.e. $\mathbf{X}'\mathbf{1} = \mathbf{0}$ with a prior of the structure

$$p(\mu, \sigma^2, \boldsymbol{\alpha}) \propto \frac{1}{\sigma^2} p(\boldsymbol{\alpha}|\sigma^2)\,. \tag{9}$$

Integrating over μ we obtain

$$\begin{aligned} p(\mathbf{y}|\sigma^2, \boldsymbol{\alpha}, \mathbf{X}) &= \int p(\mathbf{y}|\sigma^2, \mu, \boldsymbol{\alpha}, \mathbf{X}) d\mu = \\ &= \frac{1}{\sqrt{N}(2\pi\sigma^2)^{(N-1)/2}} \exp\Big(-\frac{1}{2}(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})'(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})\Big)\,, \end{aligned}$$

where $\mathbf{y}_c = \mathbf{y} - \bar{y}$. Further integration over $\boldsymbol{\alpha}$ and σ^2 yields the conditional marginal likelihood $p(\mathbf{y}|\sigma^2, \mathbf{X}) = \int p(\mathbf{y}|\sigma^2, \boldsymbol{\alpha}, \mathbf{X}) d\boldsymbol{\alpha}$ and the marginal likelihood

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\sigma^2, \mathbf{X}) \frac{1}{\sigma^2} d\sigma^2\,.$$

A.1 Conjugate Prior

Under the conjugate prior $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{a}_0, \mathbf{A}_0\sigma^2)$ analytical integration is feasible, and the conditional marginal likelihood and marginal likelihood are given as

$$p(\mathbf{y}|\sigma^2, \mathbf{X}) = \frac{1}{\sqrt{N}(2\pi\sigma^2)^{(N-1)/2}} \frac{|\mathbf{A}_N|^{1/2}}{|\mathbf{A}_0|^{1/2}} \exp\left(-\frac{S_N}{\sigma^2}\right) \tag{10}$$

$$p(\mathbf{y}|\mathbf{X}) = \frac{1}{\sqrt{N}(2\pi)^{(N-1)/2}} \frac{|\mathbf{A}_N|^{1/2}}{|\mathbf{A}_0|^{1/2}} \frac{\Gamma(s_N)}{S_N^{s_N}}\,. \tag{11}$$

Here $\mathbf{a}_N, \mathbf{A}_N$ are the moments of the posterior distribution $p(\boldsymbol{\alpha}|\sigma^2, \mathbf{y})$:

$$\mathbf{A}_N = \left(\mathbf{X}'\mathbf{X} + \mathbf{A}_0^{-1}\right)^{-1}\,, \qquad \mathbf{a}_N = \mathbf{A}_N\left(\mathbf{X}'\mathbf{y}_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right)\,,$$

and

$$S_N = \frac{1}{2}\left(\mathbf{y}_c'\mathbf{y}_c + \mathbf{a}_0'\mathbf{A}_0^{-1}\mathbf{a}_0 - \mathbf{a}_N'\mathbf{A}_N^{-1}\mathbf{a}_N\right)\,, \qquad s_N = \frac{N-1}{2}\,.$$

Special cases are the independence prior $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, c\mathbf{I})$ and the g-prior $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, g(\mathbf{X}'\mathbf{X})^{-1})$. In both cases $\mathbf{a}_N = \mathbf{A}_N\mathbf{X}'\mathbf{y}_c$ and hence S_N simplifies to

$$S_N = \frac{1}{2}\left(\mathbf{y}_c'\mathbf{y}_c - \mathbf{a}_N'\mathbf{A}_N^{-1}\mathbf{a}_N\right) = \frac{1}{2}\left(\mathbf{y}_c'\mathbf{y}_c - \mathbf{y}_c'\mathbf{X}\mathbf{A}_N\mathbf{X}'\mathbf{y}_c\right)\,.$$

For the independence prior, $|\mathbf{A}_0| = c^d$ and $\mathbf{A}_N = (\mathbf{X}'\mathbf{X} + \frac{1}{c}\mathbf{I})^{-1}$; for the g-prior $\mathbf{A}_N = \frac{g}{g+1}(\mathbf{X}'\mathbf{X})^{-1}$ and hence $|\mathbf{A}_N|^{1/2}/|\mathbf{A}_0|^{1/2} = (1+g)^{-d/2}$.

A.2 Fractional Prior

The fractional prior is obtained as a fraction of the likelihood, more specific we define the fractional prior as

$$p(\boldsymbol{\alpha}|\sigma^2) \propto p\left(\mathbf{y}_c|\boldsymbol{\alpha}, \sigma^2\right)^b \propto \exp\Big(-\frac{b}{2\sigma^2}(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})'(\mathbf{y}_c - \mathbf{X}\boldsymbol{\alpha})\Big)\,.$$

The posterior, obtained by combining the prior with the remaining fraction of the likelihood, is the normal distribution with moments

$$\mathbf{A}_N = (\mathbf{X}'\mathbf{X})^{-1}\,, \qquad \mathbf{a}_N = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_c\,.$$

Conditional marginal likelihood and marginal likelihood can be computed from formulas (10) and (11) with $S_N = \mathbf{y}_c'(\mathbf{I} - (1-b)\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})\mathbf{y}_c$ and $|\mathbf{A}_N|^{1/2}/|\mathbf{A}_0|^{1/2} = b^{d/2}$.

B Posterior Inclusion Probabilities

We compute posterior inclusion probabilities for a Dirac spike combined with i-, g- and f-slab. Without loss of generality, we consider posterior inclusion of last regressor $\mathbf{x}_d$

conditional on $\boldsymbol{\delta}_{\setminus d}$. Further, we condition on σ^2 and compute the posterior inclusion probability as

$$p(\delta_d = 1|\mathbf{y}, \boldsymbol{\delta}_{\setminus d}, \sigma^2) = \frac{1}{1 + \dfrac{p(\mathbf{y}|\boldsymbol{\delta}_{\setminus d}, \delta_d = 0, \sigma^2)}{p(\mathbf{y}|\delta_{\setminus d}, \delta_d = 1, \sigma^2)} \dfrac{(1-\omega)}{\omega}} .$$

We use the notation $\mathbf{x}_j'\mathbf{x}_j = Ns_j^2$, $\mathbf{y}_c'\mathbf{x}_j = Ns_j s_y r_{yj}$, $j = 1, \ldots, d$ and $\mathbf{y}_c'\mathbf{y}_c = Ns_y^2$ and denote by $\hat{\alpha}_j = s_{yj}/s_j^2 = r_{yj}s_y/s_j$ the LS-estimator of α_j. It will turn out that the conditional posterior inclusion probability of regressor $\mathbf{x}_d$ can be written as a function of $\hat{\alpha}_d$ and additional parameters θ, depending on the slab, as

$$p(\delta_d = 1|\mathbf{y}, \boldsymbol{\delta}_{\setminus d}, \sigma^2) = \frac{1}{1 + \exp(h(\hat{\alpha}_d, \theta)/2) \dfrac{(1-\omega)}{\omega}} .$$

B.1 Orthogonal Regressors

Let $\boldsymbol{\delta}^* = (\boldsymbol{\delta}_{\setminus d}, 1)$. For orthogonal regressors, both prior and posterior covariance matrix $\mathbf{A}_{0,\boldsymbol{\delta}^*}$ and $\mathbf{A}_{\boldsymbol{\delta}^*}$ are diagonal matrices for any of the priors on $\boldsymbol{\alpha}_{\boldsymbol{\delta}^*}$ considered here. Denoting by $\mathbf{A}_{\boldsymbol{\delta}^*,0}(d)$, $\mathbf{A}_{\boldsymbol{\delta}^*}(d)$, $\mathbf{a}_{0,\boldsymbol{\delta}^*}(d)$ and $\mathbf{a}_{\boldsymbol{\delta}^*}(d)$ the d-th element of $\mathbf{A}_{0,\boldsymbol{\delta}^*}$, $\mathbf{A}_{\boldsymbol{\delta}^*}$, $\mathbf{a}_{0,\boldsymbol{\delta}^*}$ and $\mathbf{a}_{\boldsymbol{\delta}^*}$, respectively, we obtain

$$h(\hat{\alpha}_d, \theta) = 2 \log \frac{p(\mathbf{y}|\boldsymbol{\delta}_{\setminus d}, \delta_d = 0, \sigma^2)}{p(\mathbf{y}|\boldsymbol{\delta}_{\setminus d}, \delta_d = 1, \sigma^2)} \tag{12}$$

$$= -\frac{1}{\sigma^2} \left(\frac{\left(\mathbf{a}_{\boldsymbol{\delta}^*}(d)\right)^2}{\mathbf{A}_{\boldsymbol{\delta}^*}(d)} - \frac{\left(\mathbf{a}_{0,\boldsymbol{\delta}^*}(d)\right)^2}{\mathbf{A}_{0,\boldsymbol{\delta}^*}(d)} \right) + \log \frac{\mathbf{A}_{0,\boldsymbol{\delta}^*}(d)}{\mathbf{A}_{\boldsymbol{\delta}^*}(d)} . \tag{13}$$

Further, under any of the three slabs,

$$\frac{\left(\mathbf{a}_{\boldsymbol{\delta}^*}(d)\right)^2}{\mathbf{A}_{\boldsymbol{\delta}^*}(d)} = \frac{(\mathbf{y}_c'\mathbf{x}_d)^2}{1/\mathbf{A}_{\boldsymbol{\delta}^*}(d)} = \frac{(Ns_d s_y r_{yd})^2}{1/\mathbf{A}_{\boldsymbol{\delta}^*}(d)} = \frac{(Ns_d)^2 s_d^2 \hat{\alpha}_d^2}{1/\mathbf{A}_{\boldsymbol{\delta}^*}(d)} .$$

For the i-slab with $\mathbf{a}_{0,\boldsymbol{\delta}^*}(d) = 0$, $\mathbf{A}_{0,\boldsymbol{\delta}^*}(d) = c$ and $1/\mathbf{A}_{\boldsymbol{\delta}^*}(d) = \mathbf{x}_d'\mathbf{x}_d + 1/c = Ns_d^2 + 1/c$ we get

$$(\mathbf{a}_{\boldsymbol{\delta}^*}(d))^2/\mathbf{A}_{\boldsymbol{\delta}^*} = N\hat{\alpha}_d^2 s_d^2 \frac{1}{1 + 1/(Ns_d^2 c)} .$$

Thus, h is a function of $\hat{\alpha}_d$ and c, given as

$$h(\hat{\alpha}_d, c) = -N\frac{\hat{\alpha}_d^2 s_d^2}{\sigma^2} \frac{1}{1 + 1/(Ns_d^2 c)} + \log(Ns_d^2 c + 1) .$$

For the g-slab, inserting $\mathbf{a}_{0,\boldsymbol{\delta}^*}(d) = 0$, $\mathbf{A}_{0,\boldsymbol{\delta}^*}(d) = g/(Ns_d^2)$ and $1/\mathbf{A}_{\boldsymbol{\delta}^*}(d) = (1 + 1/g)Ns_d^2$ in formula (13) yields

$$h(\hat{\alpha}_d, g) = -\frac{N\hat{\alpha}_d^2 s_d^2}{\sigma^2} \frac{1}{1 + 1/g} + \log(1 + g) .$$

Finally, as for the f-slab $\mathbf{a}_{0,\boldsymbol{\delta}^*}(d) = b\mathbf{a}_{\boldsymbol{\delta}^*}(d)$, $\mathbf{A}_{0,\boldsymbol{\delta}^*}(d) = 1/(bNs_d^2)$ and $1/\mathbf{A}_{\boldsymbol{\delta}^*}(d) = Ns_d^2$, we have

$$\frac{\left(\mathbf{a}_{\boldsymbol{\delta}^*}(d)\right)^2}{\mathbf{A}_{\boldsymbol{\delta}^*}(d)} - \frac{\left(\mathbf{a}_{0,\boldsymbol{\delta}^*}(d)\right)^2}{\mathbf{A}_{0,\boldsymbol{\delta}^*}(d)} = (1-b)N\hat{\alpha}_d^2 s_d^2$$

and hence

$$h(\hat{\alpha}_d, b) = -(1-b)\frac{N\hat{\alpha}_d^2 s_d^2}{\sigma^2} - \log(b)\,.$$

B.2 Correlated Regressors

We assume $s_j^2 = 1$, $j = 1, 2$. To compute the posterior inclusion probability of $\mathbf{x}_2$ when $\mathbf{x}_1$ is included in the model we compare the conditional marginal likelihoods of the two models $\boldsymbol{\delta} = (1, 1)$ and $\boldsymbol{\delta}^* = (1, 0)$ by

$$\begin{aligned} 2\log\frac{p(\mathbf{y}|\boldsymbol{\delta}^*)}{p(\mathbf{y}|\boldsymbol{\delta})} &= -\frac{1}{\sigma^2}\left(\mathbf{a}'_{0,\boldsymbol{\delta}^*}\mathbf{A}^{-1}_{0,\boldsymbol{\delta}^*}\mathbf{a}_{0,\boldsymbol{\delta}^*} - \mathbf{a}'_{\boldsymbol{\delta}^*}\mathbf{A}^{-1}_{\boldsymbol{\delta}^*}\mathbf{a}_{\boldsymbol{\delta}^*} - \mathbf{a}'_{0,\boldsymbol{\delta}}\mathbf{A}^{-1}_{0,\boldsymbol{\delta}}\mathbf{a}_{0,\boldsymbol{\delta}} + \mathbf{a}'_{\boldsymbol{\delta}}\mathbf{A}^{-1}_{\boldsymbol{\delta}}\mathbf{a}_{\boldsymbol{\delta}}\right) \\ &\quad + \log\frac{|\mathbf{A}_{0,\boldsymbol{\delta}}||\mathbf{A}_{\boldsymbol{\delta}^*}|}{|\mathbf{A}_{0,\boldsymbol{\delta}^*}||\mathbf{A}_{\boldsymbol{\delta}}|}\,. \end{aligned}$$

This simplifies as follows:

$$\begin{aligned} \text{i-slab:}\quad & 2\log\frac{p(\mathbf{y}|\boldsymbol{\delta}^*)}{p(\mathbf{y}|\boldsymbol{\delta})} = -\frac{1}{\sigma^2}\left(\mathbf{a}'_{\boldsymbol{\delta}}\mathbf{A}^{-1}_{\boldsymbol{\delta}}\mathbf{a}_{\boldsymbol{\delta}} - \mathbf{a}'_{\boldsymbol{\delta}^*}\mathbf{A}^{-1}_{\boldsymbol{\delta}^*}\mathbf{a}_{\boldsymbol{\delta}^*}\right) + \log\frac{c|\mathbf{A}_{\boldsymbol{\delta}^*}|}{|\mathbf{A}_{\boldsymbol{\delta}}|} \\ \text{g-slab:}\quad & 2\log\frac{p(\mathbf{y}|\boldsymbol{\delta}^*)}{p(\mathbf{y}|\boldsymbol{\delta})} = -\frac{1}{\sigma^2}\left(\mathbf{a}'_{\boldsymbol{\delta}}\mathbf{A}^{-1}_{\boldsymbol{\delta}}\mathbf{a}_{\boldsymbol{\delta}} - \mathbf{a}'_{\boldsymbol{\delta}^*}\mathbf{A}^{-1}_{\boldsymbol{\delta}^*}\mathbf{a}_{\boldsymbol{\delta}^*}\right) + \log(g+1) \\ \text{f-slab:}\quad & 2\log\frac{p(\mathbf{y}|\boldsymbol{\delta}^*)}{p(\mathbf{y}|\boldsymbol{\delta})} = -\frac{1}{\sigma^2}\left(\mathbf{a}'_{\boldsymbol{\delta}}\mathbf{A}^{-1}_{\boldsymbol{\delta}}\mathbf{a}_{\boldsymbol{\delta}} - \mathbf{a}'_{\boldsymbol{\delta}^*}\mathbf{A}^{-1}_{\boldsymbol{\delta}^*}\mathbf{a}_{\boldsymbol{\delta}^*}\right)(1-b) + \log(b)\,. \end{aligned}$$

We give details for the g-slab. Note that using the notation introduced in Section 5,

$$\mathbf{X}'\mathbf{X} = N\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{X}'\mathbf{y}_c = Ns_y\begin{pmatrix} r_{y1} \\ r_{y1} \end{pmatrix}\,.$$

$\boldsymbol{\delta}^*$ denotes the model with $\mathbf{x}_1$ as the only regressor, hence we get (as for orthogonal regressors)

$$\mathbf{a}'_{\boldsymbol{\delta}^*}\mathbf{A}^{-1}_{\boldsymbol{\delta}^*}\mathbf{a}_{\boldsymbol{\delta}^*} = \frac{g}{g+1}Nr_{y1}^2 s_y^2\,.$$

As the corresponding term for model $\boldsymbol{\delta}$ is given as

$$\mathbf{a}'_{\boldsymbol{\delta}}\mathbf{A}^{-1}_{\boldsymbol{\delta}}\mathbf{a}_{\boldsymbol{\delta}} = \frac{g}{g+1}\mathbf{y}_c\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}_c = \frac{g}{g+1}\frac{Ns_y^2}{1-r_{12}^2}\left(r_{y1}^2 - 2r_{12}r_{y1}r_{y2} + r_{y2}^2\right)\,,$$

we get

$$\mathbf{a}'_{\boldsymbol{\delta}}\mathbf{A}^{-1}_{\boldsymbol{\delta}}\mathbf{a}_{\boldsymbol{\delta}} - \mathbf{a}'_{\boldsymbol{\delta}^*}\mathbf{A}^{-1}_{\boldsymbol{\delta}^*}\mathbf{a}_{\boldsymbol{\delta}^*} = \frac{g}{g+1}\frac{Ns_y^2(r_{y2} - r_{y1}r_{12})^2}{(1-r_{12}^2)}\,,$$

and finally, using $\hat{\alpha}_2 = \frac{s_y(r_{y2} - r_{12}r_{y1})}{(1-r_{12}^2)}$, we obtain

$$2\log\frac{p(\mathbf{y}|\boldsymbol{\delta}^*)}{p(\mathbf{y}|\boldsymbol{\delta})} = -\frac{N\hat{\alpha}_2^2}{\sigma^2}(1-r_{12}^2)\,\frac{g}{g+1} + \log(g+1)\,.$$

Obviously for the f-slab we have

$$2\log\frac{p(\mathbf{y}|\boldsymbol{\delta}^*)}{p(\mathbf{y}|\boldsymbol{\delta})} = -\frac{N\hat{\alpha}_2^2}{\sigma^2}(1-r_{12}^2)(1-b) + \log(b)\,.$$

References

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

Dey, T., Ishwaran, H., and Rao, S. J. (2008). An in-depth look at highest posterior model selection. *Econometric Theory*, *24*, 377-403.

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

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*, 1360-1383.

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

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

Geweke, J. (1996). Variable selection and model comparison in regression. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. Smith (Eds.), *Bayesian Statistics 5 – Proceedings of the fifth Valencia International Meeting* (p. 609-620). Oxford University Press.

Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, *7*, 473-511.

Ishwaran, H., and Rao, S. J. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association*, *98*, 438-455.

Ishwaran, H., and Rao, S. J. (2005). Spike and slab variable selection; frequentist and Bayesian strategies. *Annals of Statistics*, *33*, 730-773.

Johnson, V. E., and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society, Series B*, *72*, 143-170.

Konrath, S., Kneib, T., and Fahrmeir, L. (2008). *Bayesian Regularisation in Structured Additive Regression Models for Survival Data* (Tech. Rep. No. 35). University of Munich, Department of Statistics.

Malsiner-Walli, G. (2010). *Bayesian Variable Selection in Normal Regression Models*. Unpublished master's thesis, Johannes Kepler Universität Linz, Institut für Angewandte Statistik.

Mitchell, T., and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *404*, 1023-1032.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B*, *57*, 99-118.

Smith, M., and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, *75*, 317-343.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267-288.

Wagner, H., and Duller, C. (2011). Bayesian model selection for logistic regression models with random intercept. *Computational Statistics and Data Analysis*. (doi:10.1016/j.csda.2011.06.033)

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques* (p. 233-243). Elsevier Sccience Publishers.

Authors' address:

Gertraud Malsiner-Walli and Helga Wagner
Department of Applied Statistics
Johannes Kepler University Linz
Freistädter Straße 315
4040 Linz
Austria

E-Mail: `Helga.Wagner@jku.at`