Bayesian One-Sided Variable Selection

Xin Gu, Herbert Hoijtink & Joris Mulder

To cite this article: Xin Gu, Herbert Hoijtink & Joris Mulder (2022) Bayesian One-Sided Variable Selection, Multivariate Behavioral Research, 57:2-3, 264-278, DOI: 10.1080/00273171.2020.1813067

To link to this article: https://doi.org/10.1080/00273171.2020.1813067

Published online: 01 Sep 2020.

MULTIVARIATE BEHAVIORAL RESEARCH
2022, VOL. 57, NOS. 2–3, 264–278
https://doi.org/10.1080/00273171.2020.1813067

Bayesian One-Sided Variable Selection

Xin Gu[a], Herbert Hoijtink[b], and Joris Mulder[c,d]

[a]Department of Educational Psychology, East China Normal University; [b]Department of Methodology and Statistics, Utrecht University; [c]Department of Methodology and Statistics, Tilburg University; [d]Jheronimus Academy of Data Science

ABSTRACT

This paper presents a novel Bayesian variable selection approach that accounts for the sign of the regression coefficients based on multivariate one-sided tests. We propose a truncated *g* prior to specify a prior distribution of coefficients with anticipated signs in a given model. Informative priors for the direction of the effects can be incorporated into prior model probabilities. The best subset of variables is selected by comparing the posterior probabilities of the possible models. The new Bayesian one-sided variable selection procedure has higher chance to include relevant variables and therefore select the best model, if the anticipated direction is accurate. For a large number of candidate variables, we present an adaptation of a Bayesian model search method for the one-sided variable selection problem to ensure fast computation. In addition, a fully Bayesian approach is used to adjust the prior inclusion probability of each one-sided model to correct for multiplicity. The performance of the proposed method is investigated using several simulation studies and two real data examples.

KEYWORDS
Fully Bayesian approach;
one-sided variable selection;
prior model probabilities;
MCMC model search;
truncated *g* prior

Introduction

Variable selection is an important step when analyzing data collected for social and behavioral studies. In regression models the objective of variable selection is to identify relevant predictor variables of an outcome variable. Irrelevant or redundant predictors should be removed before conducting regression analyses as they will add noise when estimating or testing other quantities that researchers are interested in. Statisticians have proposed many variable selection methods for regression analysis. Some modern methods include the least absolute shrinkage and selection operator (Lasso) presented by Tibshirani (1996) in the frequentist setting, variable selection using ”spike and slab” priors popularized by George and McCulloch (1993) from a Bayesian perspective, and their mixture Bayesian Lasso developed by Park and Casella (2008) and Ročková and George (2018). For an overview of shrinkage priors for Bayesian variable selection see, for instance, van Erp et al. (2019).

Variable selection is traditionally a problem of multiple two-sided hypothesis tests with respect to regression coefficients. It is well known that the one-sided test has higher power than the two-sided test. Because of this property, it has been extensively discussed, see for example, Jeffreys (1961, p. 283), Berger and Mortera (1999), and Marsman and Wagenmakers (2017). In the context of variable selection, higher power implies larger probability to include a variable when the anticipated direction of the effect is correct. While there is also a larger probability to exclude a variable if it is truly null because the null will receive more support when the observed effect is in the opposite direction. This suggests that with one-sided models, the probability of selecting the best model increases, that is, relevant variables will be included with higher probability while irrelevant variables will still be excluded. For this reason, previous studies have adopted the one-sided tests in the variable selection process. For example, Wolak (1987) first used multivariate one-sided tests to select associated variables in the regression model, and Hughes and King (2003) proposed one-sided AIC for model selection. More recently, Tibshirani et al. (2016) discussed both one-sided and two-sided tests in their variable selection procedure, and concluded that the one-sided test is preferred because it has stronger power than the two-sided test.

It is a waste of information not to include some prior information about the direction of the effects if available, as very often, even in variable selection

CONTACT Xin Gu guxin57@hotmail.com Department of Educational Psychology, East China Normal University, 3663 N. Zhongshan Rd., Shanghai 200062, China.

problems with many potential variables, researchers have expectations of the direction based on external knowledge or published work. For example, when one is interested in predicting test scores of students based on various potentially important predictor variables, such as the number of teachers in a school, the number of available computers for students to practice, the expenditure per student, etc., it is likely that the effects of these variables, if they are present, are positive on test scores of students (Stock & Watson, 2015). This paper presents a novel Bayesian algorithm for one-sided variable selection problems where researchers can incorporate prior knowledge about the direction of the effects.

In the Bayesian one-sided variable selection algorithm, the best subset of variables along with their effect directions will be selected by means of posterior model probabilities. The posterior model probability is proportional to the marginal likelihood of the data for a particular model times the prior model probability. Depending on the amount of candidate variables an algorithm with fixed prior probabilities for the one-sided models and a fully Bayesian algorithm are proposed. The latter is particularly useful in the case of many candidate variables (Scott & Berger, 2010). Furthermore, as it becomes computationally infeasible to derive the exact posterior model probabilities for all models given a large number of candidate variables, a model search method is needed to obtain numerical estimates for the one-sided models. Various model search methods using Markov chain Monte Carlo (MCMC) samples have been developed by George and McCulloch (1993), George and McCulloch (1997), Kuo and Mallick (1998), and Dellaportas et al. (2002) among others. This method can also be used to select the median probability model (Barbieri & Berger, 2004). More recently, an EM variable selection algorithm was presented as an alternative to MCMC model search methods Ročková and George (2014). In this paper, we extend the model search algorithm proposed by George and McCulloch (1997) to the one-sided variable selection problem. The algorithm will not visit all models but only those having relatively high posterior probabilities. This will substantially reduce the computation task because in practice most candidate models have almost zero posterior probability and will not be visited in the algorithm.

This paper is organized as follows. Section 2 proposes a one-sided variable selection scheme. Section 3 presents how the one-sided models can be compared by means of the Bayes factor. Thereafter, we explain the prior probability specification for the one-sided model using researchers' prior beliefs of the effect directions in Section 4. Furthermore, the fully Bayesian approach is adopted such that multiplicity can be controlled. In the case of a large number of candidate variables, Section 5 provides the MCMC model search algorithm counterpart of the one-sided selection. In Section 6, we conduct several simulation studies to investigate the performance of one-sided variable selection, as well as the MCMC model search method and the fully Bayesian approach. Subsequently, two empirical data examples are used to illustrate how the proposed variable selection scheme can be used in Section 7. This paper ends with a conclusion.

Variable selection schemes

In this paper, we consider a variable selection problem in the context of normal linear regression models:

$$y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_J x_{Ji} + \epsilon_i, \tag{1}$$

where y_i is the outcome variable, α is the intercept, $x_{1i}, \ldots, x_{Ji}$ are candidate variables with $\beta_1, \ldots, \beta_J$ being the corresponding coefficients and J the number of variables, and $\epsilon_i\ \tilde{}N(0\sigma^2)$ are the residuals with σ^2 being their variance. Our target is to select a set of relevant variables and remove others, or more specifically to identify whether each coefficient β_j for $j = 1, \ldots, J$ is equal to zero. Therefore, variable selection in regression models can be seen as a multiple test or selection problem on the regression coefficients. For example, if two variables are under consideration the problem comes down to the selection of the following models:

$$\begin{aligned} &M_0 : \beta_1 = 0; \beta_2 = 0, \quad M_1 : \beta_1 = 0; \beta_2 \neq 0, \\ &M_2 : \beta_1 \neq 0; \beta_2 = 0, \quad M_F : \beta_1 \neq 0; \beta_2 \neq 0, \end{aligned} \tag{2}$$

where M_0 and M_F denote the null and full models, respectively. For each coefficient across models the selection is two-sided, i.e., $\beta_j = 0$ against $\beta_j \neq 0$.

An alternative variable selection approach can be obtained by replacing the two-sided model by a one-sided model. The one-sided model consists of a mixture of βs that are larger than, smaller than or equal to zero. For example, with two coefficients candidate models are:

$$\begin{aligned} &M_0 : \beta_1 = 0; \beta_2 = 0, \quad M_1 : \beta_1 = 0; \beta_2 > 0, \quad M_2 : \beta_1 = 0; \beta_2 < 0, \\ &M_3 : \beta_1 > 0; \beta_2 = 0, \quad M_4 : \beta_1 > 0; \beta_2 > 0, \quad M_5 : \beta_1 > 0; \beta_2 < 0, \\ &M_6 : \beta_1 < 0; \beta_2 = 0, \quad M_7 : \beta_1 < 0; \beta_2 > 0, \quad M_8 : \beta_1 < 0; \beta_2 < 0. \end{aligned} \tag{3}$$

We refer to this as one-sided variable selection. A straightforward advantage of this approach is that besides selecting a set of variables that best predicts the outcome variable, the direction of effects of included variables can also be inferred from the results. This implies that researchers gain knowledge of whether a variable is included as well as whether the included variable has positive or negative effect on the outcome variable and with how much certainty determined by the posterior probability of the selected model. Another advantage of a Bayesian one-sided variable selection scheme is that it allows researchers to specify prior probabilities of the directions based on their prior beliefs in a direct manner, which will be discussed in the section entitled Prior model probabilities.

Bayesian variable selection

A common strategy in Bayesian variable selection is involving an indicator variable (George & McCulloch, 1993), often denoted by $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_J)$, where $\gamma_j = 1$ implies $\beta_j \neq 0$ and presence of variable x_{ji}, and $\gamma_j = 0$ implies $\beta_j = 0$ and absence of x_{ji} in the regression model. Since each γ_j is either 1 or 0, there are 2^J candidate models each of which can be denoted by M_γ. Under model M_γ, the density of the regression model is given by:

$$f(\boldsymbol{Y}|\alpha, \boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma}) = N(\alpha\boldsymbol{1} + \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, \sigma^2\boldsymbol{I}) \tag{4}$$

where $\boldsymbol{Y} = (y_1, ..., y_n)$ is an $n \times 1$ vector, $\boldsymbol{1}$ is a vector of 1 of length n, $\boldsymbol{X}_\gamma$ is an $n \times m_\gamma$ matrix with m_γ being the number of included variables, $\boldsymbol{\beta}_\gamma$ is a vector of non-zero β_j, and $\boldsymbol{I}$ denotes an $n \times n$ identity matrix.

Bayes factors for two-sided tests

The Bayes factor is a criterion when comparing two hypotheses or models, and thus can be used to select relevant variables. For example, consider the set of models in (2), we can specify either the null model M_0 or the full model M_F as the base, and compute Bayes factors of candidate models against the base. This paper will only consider the null base model M_0. In this case Bayes factors are the ratio of marginal likelihoods of the data under candidate models and the null model (Kass & Raftery, 1995):

$$BF_{\gamma 0} = m(\boldsymbol{Y}|M_\gamma)/m(\boldsymbol{Y}|M_0). \tag{5}$$

An important step when using the Bayes factor is to specify a prior distribution of unknown parameters. A commonly used prior in Bayesian variable selection is the g prior proposed by Zellner (1986):

$$\pi(\boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}_\gamma|\sigma^2, \boldsymbol{\gamma})\pi(\sigma^2|\boldsymbol{\gamma}) \tag{6}$$

with

$$\pi(\boldsymbol{\beta}_\gamma|\sigma^2, \boldsymbol{\gamma}) = N(\boldsymbol{0}, g\sigma^2(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma)^{-1}) \tag{7}$$

and

$$\pi(\sigma^2|\boldsymbol{\gamma}) \propto \sigma^{-2}. \tag{8}$$

Without loss of generality we assume that an intercept is included in every model. An attractive advantage of the g prior is that it is conjugate to the density $f(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma})$, and it leads to a closed form of the Bayes factor (Garcia-Donato & Martinez-Beneito, 2013; Liang et al., 2008):

$$BF_{\gamma 0} = \frac{(1+g)^{(n-1-m_\gamma)/2}}{(1+g\frac{RSS}{TSS})^{(n-1)/2}} \tag{9}$$

where $RSS = (\boldsymbol{y} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)^T(\boldsymbol{y} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)$ is the residual sum of squares with $\hat{\boldsymbol{\beta}}_\gamma$ being the OLS estimate of $\boldsymbol{\beta}_\gamma$ under M_γ, and $TSS = \boldsymbol{y}^T\boldsymbol{y}$ is the total sum of squares. The hyper-parameter g is of crucial importance when using Bayes factors based on g priors. With the increase of g the support for the null hypothesis increases (Gu et al., 2016) and therefore fewer variables will be included in the selected model. The specification of g has been thoroughly discussed in Liang et al. (2008) and Consonni et al. (2018). This paper considers two common choices: $g = n$ which corresponds to the unit information prior (Kass & Wasserman, 1995) and $g = J^2$ based on risk inflation criterion (Foster & George, 1994).

Bayes factors for one-sided tests

For one-sided variable selection, we use an indicator variable $\boldsymbol{\gamma}'$ to represent the one-sided model. Possible values of $\boldsymbol{\gamma}'$ are $\gamma'_j = \{0, 1, -1\}$ where $\grave{\gamma_j}$ equal to 0, 1 and -1 corresponds to $\beta_j = 0, \beta_j > 0$ and $\beta_j < 0$, respectively. The Bayes factor of the one-sided model $M_{\gamma'}$ against the null M_0 can be written as:

$$BF_{\gamma' 0} = BF_{\gamma'\gamma}BF_{\gamma 0}, \tag{10}$$

where $BF_{\gamma'\gamma}$ is the Bayes factor for the one-sided model $M_{\gamma'}$ against the two-sided M_γ. Using the same prior given by Equation (6) under the two-sided model, $BF_{\gamma 0}$ can be obtained through Equation (9). To compute $BF_{\gamma'\gamma}$, note that the one-sided model $M_{\gamma'}$ is nested in the two-sided model M_γ, and therefore we can use the encompassing prior method proposed by Klugkist et al. (2005) where the prior under the one-sided model is a truncation of the prior under the two-sided model, that is,

$$\pi(\boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma}') = \pi(\boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})/\Phi_{prior}, \tag{11}$$

where $\Phi_{prior} = \int_{\boldsymbol{\beta}_\gamma \in \mathcal{B}_{\gamma'}} \pi(\boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma}) d\boldsymbol{\beta}_\gamma d\sigma^2$ is the probability of prior distribution $\pi(\boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma})$ truncated in

$$\mathcal{B}_{\gamma'} = \{\boldsymbol{\beta}_\gamma | \boldsymbol{\gamma}^* \boldsymbol{\beta}_\gamma \in \mathbb{R}^{+m_\gamma}\}, \tag{12}$$

where $\boldsymbol{\gamma}^*$ is equal to $\boldsymbol{\gamma}'$ with the zero elements omitted. The element of $\boldsymbol{\beta}_\gamma \in \mathcal{B}_{\gamma'}$ is $\beta_j > 0$ if $\gamma'_j = 1$ and $\beta_j < 0$ if $\gamma'_j = -1$. Note that the prior in (11) is a truncated version of the g prior, where the prior mean is zero. This assumes that small effects (i.e., effects close to zero) are more plausible than large effects (i.e., effects far away from zero), which is a reasonable assumption in the social and behavioral sciences.

Using the encompassing prior approach, the Bayes factor for the one-sided model against the two-sided can be written as:

$$BF_{\gamma'\gamma} = \frac{\Phi_{posterior}}{\Phi_{prior}}, \tag{13}$$

where $\Phi_{posterior} = \int_{\boldsymbol{\beta}_\gamma \in \mathcal{B}_{\gamma'}} \pi(\boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma}, \boldsymbol{Y}) d\boldsymbol{\beta}_\gamma d\sigma^2$ with $\pi(\boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma}, \boldsymbol{Y})$ being the posterior distribution of $\boldsymbol{\beta}_\gamma$ and σ^2 under model M_γ. The derivation of Equation (13) can be found in Gu et al. (2018). As can be seen from Equation (13), the Bayes factor $BF_{\gamma'\gamma}$ can be expressed as the ratio of posterior and prior probabilities that the unconstrained $\boldsymbol{\beta}_\gamma$ lie in $\mathcal{B}_{\gamma'}$.

To simplify the computation of Φ_{prior} and $\Phi_{posterior}$, we treat σ^2 as known and set it equal to its least squares estimator $\hat{\sigma}^2$. This setting was also suggested by George and Foster (2000). On the one hand, the prior probability Φ_{prior} is invariant for the choice of σ^2 when evaluating the one-sided hypotheses/models, which has been demonstrated by Mulder (2014). Therefore, this simplification will not influence the value of Φ_{prior}. Given $\sigma^2 = \hat{\sigma}^2, \Phi_{prior}$ becomes cumulative probability of a normal distribution constrained in $\boldsymbol{\beta}_\gamma \in \mathcal{B}_{\gamma'}$:

$$\begin{aligned}\Phi_{prior} &\approx \int_{\boldsymbol{\beta}_\gamma \in \mathcal{B}_{\gamma'}} \pi(\boldsymbol{\beta}_\gamma|\hat{\sigma}^2, \boldsymbol{\gamma}) d\boldsymbol{\beta}_\gamma \\ &= \int_{\boldsymbol{\beta}_\gamma \in \mathcal{B}_{\gamma'}} N(\boldsymbol{0}, g\hat{\sigma}^2(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}) d\boldsymbol{\beta}_\gamma. \end{aligned} \tag{14}$$

On the other hand, by letting σ^2 be known the posterior of $\boldsymbol{\beta}_\gamma$ is approximated by a normal distribution and $\Phi_{posterior}$ can be obtained by:

$$\begin{aligned}\Phi_{posterior} &\approx \int_{\boldsymbol{\beta}_\gamma \in \mathcal{B}_{\gamma'}} \pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \boldsymbol{Y}) d\boldsymbol{\beta}_\gamma \\ &\approx \int_{\boldsymbol{\beta}_\gamma \in \mathcal{B}_{\gamma'}} N\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}}_\gamma, \frac{g}{1+g}\hat{\sigma}^2(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}\right) d\boldsymbol{\beta}_\gamma, \end{aligned} \tag{15}$$

where $\hat{\boldsymbol{\beta}}_\gamma$ is the OLS estimate of $\boldsymbol{\beta}_\gamma$.

Until now Bayes factors of a one-sided model against the null have been presented. The next step is to specify prior probabilities for the one-sided models, which will be discussed in the next section.

Prior model probabilities

This section presents two different techniques for specifying prior model probabilities for one-sided variable selection. One is based on fixed probabilities and the other uses a fully Bayesian approach. Depending on the amount of prior information, either informative priors or non-informative priors can be specified under both techniques.

Fixed prior probabilities of one-sided models

In default Bayesian two-sided variable selection, each variable has an equal chance of being included priori. Therefore, we set all variables the same inclusion probability $P(\gamma_j \neq 0) = p$ for $j = 1, ..., J$. Assuming independence of $\gamma_1, \gamma_2, ..., \gamma_J$, the prior model probability is given by

$$P(M_\gamma) = p^{m_\gamma}(1-p)^{J-m_\gamma}, \tag{16}$$

where m_γ is the number of included variables. A reasonable default specification of prior inclusion probability is to set $p = 1/2$ where a variable can either be included or excluded with equal prior probability.

For one-sided variable selection, we set equal prior inclusion probability $p = 1/2$ for all variables which is similar as the two-sided approach. However, the prior inclusion probability in the one-sided model is defined by the sum of the prior probabilities for the positive and negative effects, denoted by p_{j+} and p_{j-}, respectively, for variable x_j where $p_{j+} + p_{j-} = p$. These two probabilities can be chosen either equal in the case of no prior preference about the direction of the effects or unequal in case researchers have prior beliefs about the anticipated direction.

First, if researchers firmly believe that variable x_j, if included, has a positive effect, then a prior probability of $p_{j+} = 1/2$ is set leaving $p_{j-} = 0$. In this case, models with $\beta_j < 0$ will have a prior probability of zero and therefore be excluded. An opposite setup can be used if a negative effect is anticipated. We refer to the first case as strong prior beliefs. Second, if researchers expect that the effect of x_j is positive but are not completely certain, then the positive direction receives a larger prior probability than the negative. A possible setting would be $p_{j+} = 2p_{j-}$ leading to $p_{j+} = 1/3$ and $p_{j-} = 1/6$ given the choice of $p = 1/2$. In contrast, if the negative effect is favored then $p_{j+} = 1/6$ and $p_{j-} = 1/3$ can be set. We refer to the second case as moderate prior beliefs. Third, if researchers have little prior knowledge about the effect direction, an equal prior probability $p_{j+} = p_{j-} = 1/4$ will be given by default. We refer to the third case as no prior beliefs. Consequently, the prior probability of a one-sided model $M_{\gamma'}$ is the product of the prior probabilities of all variables having positive, negative, and no effects:

$$P(M_{\gamma'}) = \prod_{\gamma'_j=1} p_{+j} \prod_{\gamma'_j=-1} p_{-j} \prod_{\gamma'_j=0} (1-p). \tag{17}$$

For example, with strong prior belief for $\beta_1 > 0$ and moderate prior belief for $\beta_2 < 0$, the models $M_0, \ldots, M_8$ in Equation (3) will receive prior probabilities of $\left(\frac{1}{4}, \frac{1}{12}, \frac{1}{6}, \frac{1}{4}, \frac{1}{12}, \frac{1}{6}, 0, 0, 0\right)$, respectively.

Fully Bayesian approach for one-sided models

The specification of equal inclusion probabilities regardless of the number of variables J causes the algorithm to include more variables as J increases. This phenomenon, called multiplicity, arises from multiple tests or comparisons in variable selection. For example, in a simple case where candidate variables are independent and all have positive effects, the one-sided variable selection is executed as J independent tests of $\beta_j = 0$ against $\beta_j > 0$. Therefore, the choice $p = 1/2$ suggests a model size of $J/2$ a priori since each variable has a probability of 1/2 of being included and there are J variables. To control for multiplicity in Bayesian variable selection, previous studies have presented two approaches: the empirical Bayes approach (George & Foster, 2000) and the fully Bayesian approach (Scott & Berger, 2010). The empirical Bayes approach can be criticized because, in a way, it uses the data twice in the variable selection procedure. Therefore, this paper only adopts the fully Bayesian approach for multiplicity correction as the number of candidate variables J grows.

Instead of fixing $p = 1/2$, the fully Bayesian approach assigns a Beta distribution on the prior inclusion probability $\pi(p) = Be(a, b)$ with the default choice of $a = b = 1$ implying a uniform prior on p. This renders a prior probability of

$$P_F(M_\gamma) = \int_0^1 p(M_\gamma)\pi(p)dp = \frac{\Gamma(m_\gamma + 1)\Gamma(J - m_\gamma + 1)}{\Gamma(J+2)} \tag{18}$$

for the two-sided model, where $\Gamma(\cdot)$ is the gamma function, and $P_F(\cdot)$ denotes prior probabilities under the fully Bayesian approach.

The fully Bayesian approach can be extended to one-sided variable selection. Because the effect has different directions (or the effect is zero) with various degrees of prior beliefs, three situations of the effect directions are considered. First, for variables with strong prior beliefs of the effect directions, the prior inclusion probability p is equal to the prior probability of the anticipated direction as the opposite direction receives a probability of zero. Therefore, a Beta distribution $Be(1,1)$ can be specified for p for these variables, resulting in a probability of

$$P_{strong} = \frac{\Gamma(m_{\gamma_s} + 1)\Gamma(m_{0_s} + 1)}{\Gamma(m_{\gamma_s} + m_{0_s} + 2)}, \tag{19}$$

where m_{γ_s} and m_{0_s} are the numbers of included and excluded variables, respectively, with strong prior beliefs of the effect directions. Note that $m_{\gamma_s} + m_{0_s}$ is not equal to the number of candidate variables, but the number of variables with respect to which there are strong prior beliefs. Note also that models containing any variable of which the effect direction is opposite to the prior belief receive a prior probability of zero.

Second, for variables with moderate prior beliefs of the effect directions, we specify a Dirichlet distribution $Dirichlet(a_1, a_2, a_3)$ for the prior probabilities of the anticipated and opposite directions, and zero effect, where a_1, a_2 and a_3 are the corresponding parameters. A possible setting is $a_1 = 1/3, a_2 = 1/6$ and $a_3 = 1/2$ because a_1, a_2 and a_3 are proportional to the expected means of the prior probabilities. This gives a probability of

$$P_{moderate} = \frac{\Gamma\left(m_{\gamma_{m1}} + \frac{1}{3}\right)\Gamma\left(m_{\gamma_{m2}} + \frac{1}{6}\right)\Gamma\left(m_{0_m} + \frac{1}{2}\right)}{\Gamma\left(m_{\gamma_{m1}} + m_{\gamma_{m2}} + m_{0_m} + 1\right)} / c_m, \tag{20}$$

where $m_{\gamma_{m1}}$ and $m_{\gamma_{m2}}$ are the numbers of included variables of which the effect directions are in line with and opposite to the moderate prior beliefs, respectively, m_{0_m} is the number of excluded variables having moderate prior beliefs on the effect directions, and $c_m = \Gamma\left(\frac{1}{3}\right)\Gamma\left(\frac{1}{6}\right)\Gamma\left(\frac{1}{2}\right)$.

Third, for variables with no prior beliefs, both directions should receive an equal prior probability. Thus, there is no need to distinguish the effect direction when specifying the prior probability. Analogous to the two-sided approach, we specify a Beta distribution $Be(1,1)$ for the prior inclusion probability p, which leads to a probability of

$$P_{no} = \frac{\Gamma(m_{\gamma_n} + 1)\Gamma(m_{0_n} + 1)}{\Gamma(m_{\gamma_n} + m_{0_n} + 2)}, \tag{21}$$

where m_{γ_n} and m_{0_n} are the numbers of included and excluded variables, respectively, with no prior beliefs of the effect directions.

The situations discussed above cover all variables (for which there are strong, median, and no prior beliefs). Consequently, prior model probabilities under the fully Bayesian approach can be computed by

$$P_F(M_{\gamma'}) = P_{strong} P_{moderate} P_{no}. \tag{22}$$

The fully Bayesian prior favors the simpler model when the number of candidate variables is large, and therefore controls multiplicity (Scott & Berger, 2010). The performance of the fully Bayesian approach in terms of multiplicity correction will be illustrated in Section 6.2.

Using the Bayes factor obtained from Equation (13) and the prior model probability computed by Equation (17) or (22), the posterior probability of a one-sided model can be obtained by

$$P(M_{\gamma'}|\boldsymbol{Y}) \propto P(M_{\gamma'}) BF_{\gamma'0} \tag{23}$$

with the fixed prior model probability, or

$$P_F(M_{\gamma'}|\boldsymbol{Y}) \propto P_F(M_{\gamma'}) BF_{\gamma'0} \tag{24}$$

using the fully Bayesian approach. The set of variables having the largest posterior model probability will be selected.

MCMC model search method

When the number of candidate variables is large, exhaustive calculation of the posterior model probabilities in (23) or (24) for all possible models becomes infeasible. For example, given $J = 20$ there are more than three billions (3^{20}) possible models under consideration. This issue can be addressed by using a MCMC model search method to the one-sided Bayesian variable selection algorithm with different prior model probability settings.

A popular MCMC model search method was proposed by George and McCulloch (1993), where a spike and slab prior distribution is set for the coefficients $\boldsymbol{\beta}$. Given a model M_γ, if $\gamma_j = 0$ then β_j has a normal prior with a mean of zero and small variance, whereas if $\gamma_j \neq 0$ then β_j has a normal prior with a mean of zero and large variance. The two prior variances are of crucial importance when using the spike and slab method. Note that the normal prior proposed in George and McCulloch (1993) is not conjugate to the likelihood of the regression model. An alternative setup is to use a conjugate prior where $\boldsymbol{\beta}|\sigma^2$ is normally distributed. George and McCulloch (1997) thoroughly discussed the non-conjugate and conjugate spike and slab priors, and their implementation in the MCMC model search method. They concluded that the conjugate form offers the advantage of analytical simplification and more efficient exploration with more correlated designs in the model search method.

In this paper, the g prior distribution presented in Equation (6) is a special case of the conjugate spike and slab prior where the small variance is set as zero if $\gamma_j = 0$, and the large variance is set as $g\sigma^2(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}$. This prior setting has been widely used in Bayesian variable selection where the MCMC model search method is adopted when the model space is large. Examples can be found in George and Foster (2000), Liang et al. (2008), and Garcia-Donato and Martinez-Beneito (2013). The hyper-parameter g plays an important role in the variable selection. The larger the g, the fewer the variables included as the Bayes factor will favor the null hypothesis more. As was elaborated in Section 3.1, two commonly used choices are: $g = n$ and $g = J^2$. However, other reasonable choices can also be specified in the proposed algorithm.

The basic idea of the MCMC algorithm for Bayesian variable selection is to sequentially sample $\boldsymbol{\gamma}$ from its posterior distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{Y})$, and select the best model which appears most often in the sample of $\boldsymbol{\gamma}$. It is important to note that when using conjugate priors the marginal posterior distribution of $\boldsymbol{\gamma}$ has an analytical form:

$$\pi(\boldsymbol{\gamma}|\boldsymbol{Y}) = P(M_\gamma|\boldsymbol{Y}) \propto BF_{\gamma 0} P(M_\gamma), \tag{25}$$

where $BF_{\gamma 0}$ is given by Equation (9) and $P(M_\gamma)$ is given by Equation (16) or (18). Because of integrating out $\boldsymbol{\beta}$ and σ^2 in $\pi(\boldsymbol{\gamma}|\boldsymbol{Y})$, we can apply the Gibbs sampler algorithm only to $\boldsymbol{\gamma}$, i.e., to sequentially sample along γ_j^t for $j = 1, \ldots, J$ and $t = 1, \ldots, T$ with T the sample size:

$$\gamma_1^0, \ldots, \gamma_J^0, \gamma_1^1, \ldots, \gamma_J^1, \ldots, \gamma_1^t, \ldots, \gamma_J^t, \ldots, \tag{26}$$

where $\gamma_1^0, \ldots, \gamma_J^0$ denote the initial values, which can be set as zero. In the Gibbs algorithm the subsequent values of γ_j^t can be sampled from its conditional posterior distribution given the latest values of all other γs.

As was pointed out by George and McCulloch (1997), the conditional distribution of γ_j given all other γs is Bernoulli. At iteration t the probability of sampling $\gamma_j^t = 1$ is

$$P(\gamma_j^t = 1|\boldsymbol{\gamma}_{-j}^t, \boldsymbol{Y}) = \frac{\pi(\gamma_j^t = 1, \boldsymbol{\gamma}_{-j}^t|\boldsymbol{Y})}{\pi(\gamma_j^t = 1, \boldsymbol{\gamma}_{-j}^t|\boldsymbol{Y}) + \pi(\gamma_j^t = 0, \boldsymbol{\gamma}_{-j}^t|\boldsymbol{Y})}, \tag{27}$$

where $\boldsymbol{\gamma}^t_{-j} = (\gamma^t_1, \ldots, \gamma^t_{j-1}, \gamma^{t-1}_{j+1}, \ldots, \gamma^{t-1}_J)$ denotes the latest values of $\boldsymbol{\gamma}$ except γ_j. Note that when sampling $\gamma^t_j, (\gamma_{j+1}, \ldots, \gamma_J)$ have not been sampled at iteration t and thus their values at the $t-1$ iteration are used. In Equation (27), $\pi(\gamma^t_j = 1, \boldsymbol{\gamma}^t_{-j}|\boldsymbol{Y})$ and $\pi(\gamma^t_j = 0, \boldsymbol{\gamma}^t_{-j}|\boldsymbol{Y})$ can be computed using Equation (25) given $\gamma^t_j = 1$ and $\gamma^t_j = 0$ respectively, and the latest $\boldsymbol{\gamma}^t_{-j}$. The probability of sampling $\gamma^t_j = 0$ is $1 - P(\gamma^t_j = 1|\boldsymbol{\gamma}^t_{-j}, \boldsymbol{Y})$.

With the two sampling probabilities, either $\gamma^t_j = 1$ or $\gamma^t_j = 0$ will be sampled, indicating the inclusion or exclusion of variable x_j in the model at iteration t. Thereafter, the algorithm visits the next γ^t_{j+1}. Once all $\boldsymbol{\gamma}^t$ have been sampled, the algorithm goes to the $t+1$ iteration until the Gibbs chain converges, to obtain the samples shown in (26). After obtaining the Gibbs samples and discarding the burn-in phase (say the first 1000 iterations for example), the best model will be the one with the highest frequency in the useful samples.

The above algorithm can be extended to one-sided variable selection where each γ'_j has three possible values 0, 1 and -1 for zero, positive and negative effects, respectively. By using the truncated g prior the posterior distribution $\pi(\boldsymbol{\gamma}'|\boldsymbol{Y})$ has a closed form as well, which can be computed through $\pi(\boldsymbol{\gamma}'|\boldsymbol{Y}) = P(M_{\gamma'}|\boldsymbol{Y})$ in (23) or $\pi(\boldsymbol{\gamma}'|\boldsymbol{Y}) = P_F(M_{\gamma'}|\boldsymbol{Y})$ in (24). Similarly, Gibbs sampler is used to sample $\boldsymbol{\gamma}'$ from its posterior distribution. For one-sided variable selection, we generalize the Bernoulli conditional distribution (27) to a multinomial distribution. The three probabilities of sampling $\gamma'^t_j = r$ for $r = 0, 1, -1$ at iteration t are

$$P(\gamma'^t_j = r|\boldsymbol{\gamma}'^{\,t}_{-j}, \boldsymbol{Y}) = \frac{\pi(\gamma'^t_j = r, \boldsymbol{\gamma}'^{\,t}_{-j}|\boldsymbol{Y})}{\sum_r \pi(\gamma'^t_j = r, \boldsymbol{\gamma}'^{\,t}_{-j}|\boldsymbol{Y})} \tag{28}$$

where $\boldsymbol{\gamma}'^{\,t}_{-j} = (\gamma'^{\,t}_1, \ldots, \gamma'_{j-1}t, \gamma'^{\,t-1}_{j+1}, \ldots, \gamma'^{\,t-1}_J)$, and $\pi(\gamma'^t_j = r, \boldsymbol{\gamma}'^{\,t}_{-j}|\boldsymbol{Y})$ can be computed using (23) or (24) given $\gamma'^t_j = r$ and the latest $\boldsymbol{\gamma}'^{\,t}_{-j}$. Based on the conditional distribution (28), the algorithm for sampling $\boldsymbol{\gamma}'$ can be implemented as follows:

```
**Algorithm 1** Gibbs sampler for Bayesian one-sided variable selection
Initialize $\boldsymbol{\gamma}'^0 = \boldsymbol{0}$ and $t = 1$

repeat

    for $j = 1, \ldots, J$ do
        Sample $\gamma'^t_j = 0$ with probability $P(\gamma'^t_j = 0|\boldsymbol{\gamma}'^{\,t}_{-j}, \boldsymbol{Y})$.
        Sample $\gamma'^t_j = 1$ with probability $P(\gamma'^t_j = 1|\boldsymbol{\gamma}'^{\,t}_{-j}, \boldsymbol{Y})$.
        Sample $\gamma'^t_j = -1$ with probability $P(\gamma'^t_j = -1|\boldsymbol{\gamma}'^{\,t}_{-j}, \boldsymbol{Y})$.
    end for
    Set $t = t + 1$.
until Gibbs chain converges
```

Convergence of the chain will be discussed in the simulation study in Section 6.2. After obtaining a sample $\boldsymbol{\gamma}'^t$ for $t = 1, \ldots, T$, we can estimate the posterior distribution of $\boldsymbol{\gamma}'$, based on which the best model (subset of variables) that has the largest probability in the distribution will be selected.

The MCMC model search method is needed when exhaustive computation of posterior probabilities for all models is infeasible. For traditional two-sided variable selection, a full enumeration of all models usually requires the number of candidate variables $J \leq 25$, see for example George and McCulloch (1997). This implies a limitation of 2^{25} models under consideration. Given a similar limited number of possible models, the MCMC model search method for 3^J one-sided models should be used when $J > 15$.

Simulation studies

This section conducts two simulation studies to investigate the difference between the two-sided and one-sided variable selection schemes presented in Section 2, and to assess the performance of the fully Bayesian approach proposed in Section 4.2 and the MCMC model search method introduced in Section 5 in the case of a large number of candidate variables. In both simulations, the hyper-parameter in the g prior is chosen as $g = n$ or $g = J^2$.

Variable selection with fixed prior model probabilities

We consider a simple variable selection problem with $J = 8$ candidate variables $x_1, \ldots, x_8$ of length n. The first four variables are independently generated from the standard normal distribution, i.e., $x_1, \ldots, x_4$ iid $\sim N(0, 1)$, while the last four variables are generated by

$$[x_5, x_6, x_7, x_8] = \boldsymbol{E} + [x_1, x_2, x_3, x_4] \times \boldsymbol{D}, \tag{29}$$

where $[\cdot]$ denotes the data matrix of the corresponding variables, $\boldsymbol{E}$ is an $n \times 4$ matrix of the independent standard normal variates, and $\boldsymbol{D}$ is a 4×4 matrix to account for possible multicollinearity of the variables. In the simulation, two choices of $\boldsymbol{D}$ are considered. The first $\boldsymbol{D} = \boldsymbol{0}$ implies no correlation among the eight variables in the population. The second $\boldsymbol{D} = \boldsymbol{D}_c = [0.5, 0.7, 0.9, 1.1]' \times [1, 1, 1, 1]$ implies strong correlation among the last four variables and moderate correlation between the first four and the last four variables. This sampling strategy of the predictors is commonly used in the simulation study for the Bayesian variable selection, see e.g., Nott and Kohn (2005). The outcome variable y_i is calculated by:

Table 1. Inclusion proportions of variables from 1000 datasets given $n = 20$ and $b = 0.4$ with prior beliefs (i) and (ii).

		$\boldsymbol{D} = \boldsymbol{0}$					$\boldsymbol{D} = \boldsymbol{D}_c$				
		Two	One (i)		One (ii)		Two	One (i)		One (ii)	
g	x_j		$+$	–	$+$	–		$+$	–	$+$	–
$g = n$	x_1	0.434	0.580	0.000	0.000	0.009	0.390	0.572	0.000	0.000	0.009
	x_2	0.458	0.000	0.610	0.003	0.000	0.484	0.000	0.565	0.000	0.000
	x_3	0.164	0.113	0.000	0.000	0.132	0.217	0.097	0.000	0.000	0.153
	x_4	0.428	0.490	0.001	0.286	0.004	0.302	0.413	0.000	0.218	0.006
	x_5	0.408	0.000	0.474	0.001	0.278	0.438	0.001	0.479	0.013	0.243
	x_6	0.151	0.078	0.051	0.035	0.088	0.150	0.080	0.047	0.026	0.070
	x_7	0.414	0.416	0.001	0.361	0.000	0.510	0.458	0.001	0.382	0.001
	x_8	0.145	0.067	0.063	0.059	0.049	0.168	0.073	0.053	0.074	0.050
$g = J^2$	x_1	0.342	0.459	0.000	0.000	0.003	0.304	0.442	0.000	0.000	0.001
	x_2	0.348	0.000	0.476	0.002	0.000	0.402	0.000	0.472	0.000	0.000
	x_3	0.100	0.077	0.000	0.000	0.072	0.140	0.067	0.000	0.000	0.080
	x_4	0.335	0.384	0.001	0.212	0.000	0.244	0.318	0.001	0.167	0.003
	x_5	0.304	0.001	0.361	0.001	0.201	0.320	0.001	0.358	0.007	0.172
	x_6	0.101	0.049	0.032	0.020	0.057	0.100	0.058	0.030	0.013	0.041
	x_7	0.329	0.338	0.000	0.271	0.000	0.404	0.370	0.001	0.277	0.000
	x_8	0.089	0.045	0.040	0.037	0.034	0.111	0.054	0.034	0.044	0.025

The results for the two-sided approach are shown under "Two." The results for the one-sided approach with prior beliefs (i) and (ii) are shown under "One (i)" and "One (ii)," respectively. The proportions of selecting variables with positive and negative effects are shown under "+" and "–," respectively.

$$y_i = \beta_1 x_{1i} + \ldots + \beta_8 x_{8i} + \epsilon_i, \qquad (30)$$

where $\epsilon_i \sim N(0, 1)$. We assume an intercept of zero in the simulation. The true values of the coefficients are given by $\boldsymbol{\beta} = (b, -b, 0, b, -b, 0, b, 0)$ where b is varied from 0 to 0.8. Based on Equation (30) and the true $\boldsymbol{\beta}$, data x_i and y_i are generated 1000 times. For each dataset, the posterior probabilities of different models will be computed and the best model with the highest posterior probability will be selected under both the two-sided and one-sided approaches.

In this simulation, the prior inclusion probability is set as $p = 1/2$ based on fixed prior model probabilities. For one-sided variable selection, we consider four scenarios for the prior beliefs of the effect directions of the eight variables:

i. prior belief: $(s+, s-, s+, m+, m-, m+, n\pm, n\pm)$
ii. prior belief: $(s-, s+, s-, m-, m+, m-, n\pm, n\pm)$
iii. prior belief: $(s+, s-, n\pm, s+, s-, n\pm, s+, n\pm)$
iv. prior belief: $(m+, m-, n\pm, m+, m-, n\pm, m+, n\pm)$

where $s+$ and $s-$ denote strong prior beliefs for the positive and negative effects, respectively, $m+$ and $m-$ denote moderate prior beliefs for the positive and negative effects, respectively, and $n\pm$ denotes no prior belief. Note that the anticipated signs of β_1, β_2, β_4 and β_5 in prior beliefs (i), (iii) and (iv) are in agreement with the signs of the true coefficients whereas the prior belief (ii) specifies opposite signs to the true. Since there are only eight candidate variables, we need not use the MCMC model search method in this simulation.

First of all, Table 1 displays the inclusion proportion of each variable from 1000 samples given $n = 20$ and $b = 0.4$ under $\boldsymbol{D} = \boldsymbol{0}$ and $\boldsymbol{D} = \boldsymbol{D}_c$ using both two-sided and one-sided selection approaches with prior beliefs (i) and (ii) based on $g = n$ and $g = J^2$. For the one-sided approach, the proportions of selecting variables with positive and negative effects are shown under the columns named by "+" and "–", respectively. The inclusion proportion is the sum of the values under "+" and "–". It can be seen from Table 1 that the proportions of including variables x_1, x_2, x_4 and x_5 under the one-sided approach with prior belief (i) are always larger than those under the two-sided approach. For example, given $\boldsymbol{D} = \boldsymbol{0}$ and $g = n$ the proportion of including x_2 is 0.458 under the two-sided approach and 0.610 under the one-sided approach with strong prior belief for the negative effect. This implies that the one-sided variable selection which involves correct prior knowledge of the effect direction can increase the probability of including variables that have actual effects. If the prior belief is opposite to the true direction, as shown in the results below "One (ii)" the inclusion proportions under the one-sided approach are smaller than those under the two-sided for the four variables. However, the one-sided approach still makes the correct direction more often for x_4 and x_5 with moderate prior beliefs of the wrong directions. It is also important to note that for variables x_3, x_6 and x_8 that have no effect in the population, the one-sided approach produces smaller inclusion proportion than the two-sided. This indicates that the one-sided variable selection does not increase the chance of incorrect inclusion compared to the two-sided approach.

Next, we will explore how often the true model is selected when the sample size is varied from $n = 10$ to

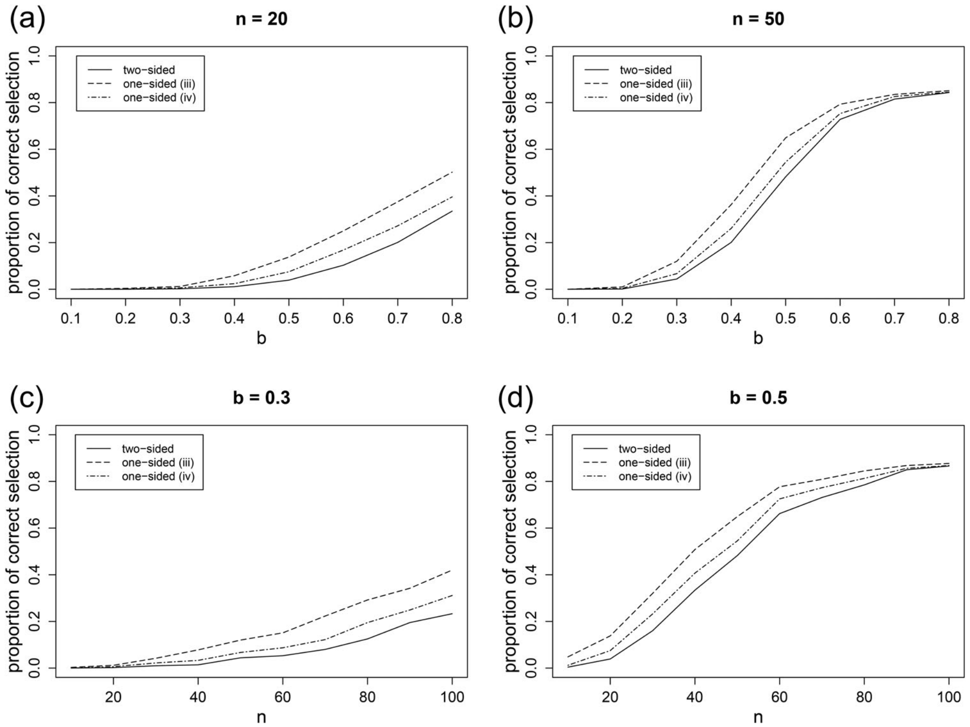

Figure 1. Proportion of selecting the true model given $\boldsymbol{D} = \boldsymbol{0}$ and $g = n$ using both two-sided and one-sided variable selection approaches. (a) and (b) set $n = 20$ and $n = 50$, respectively, and b varied from $b = 0.1$ to $b = 0.8$; (c) and (d) set $b = 0.3$ and $b = 0.5$, respectively, and n varied from $n = 10$ to $n = 100$.

$n = 100$ and the effect size is varied from $b = 0.1$ to $b = 0.8$ given $g = n$ and $\boldsymbol{D} = \boldsymbol{0}$. We only consider $\boldsymbol{D} = \boldsymbol{0}$ for independent variables because in this case the true model is clearly defined which includes variables x_1, x_2, x_4, x_5 and x_7, and excludes others. Figure 1 plots the probability of selecting the true model from the 1000 samples as a function of b given $n = 20$ and $n = 50$ (top panels), as well as a function of n given $b = 0.3$ and $b = 0.5$ (bottom panels) under the two selection approaches. The one-sided approach involves strong prior belief (iii) and moderate prior belief (iv) of the effect directions for variables that have actual effects. Note that the one-sided approach also requires correct selection of the effect directions. As can be seen from all four figures, the probability of correction selection increases as n and/or b grows up, which implies that the larger the true effect or the sample size, the higher the chance to select the true model. More interestingly, the proportion of correct selection under the one-sided approach is always higher than that under the two-sided. For example, given $n = 20$ and $b = 0.5$ in Figure 1(a) the probabilities of correct selection are 0.138 and 0.075 under the one-sided approach with prior beliefs (iii) and (iv), respectively, which are about three and two times larger than the corresponding probability of 0.039 under the two-sided approach. This indicates that the one-sided approach with correct prior beliefs of the direction performs better. On the other hand, for relatively large effects or sample sizes, Figure 1(b,d) demonstrate that the one-sided and two-sided approaches become similar and both result in large probabilities to select the true model given, for example, $n = 50$ and $b = 0.8$ in Figure 1(b) or $b = 0.5$ and $n = 100$ in Figure 1(d).

Performance of the model search method and fully Bayesian approach

In this subsection, a simulation study is conducted to evaluate the performance of the MCMC model search method and the fully Bayesian approach for one-sided variable selection. We consider various numbers of candidate variables from $J = 6$ to $J = 60$. Candidate variables $x_{1i}, \ldots, x_{Ji}$ with sample size $n = 100$ are simulated independently from the standard normal distribution $N(0,\ 1)$, and the outcome variable y_i is calculated based on Equation (1) with intercept $\alpha = 0$, residual variance $\sigma^2 = 1$, and true coefficients of

$$\beta_1 = \beta_2 = 0.5; \beta_3 = \beta_4 = 0.3; \beta_5 = \cdots = \beta_J = 0 \tag{31}$$

We assume strong, moderate and no prior beliefs of the positive direction for variables that have effects

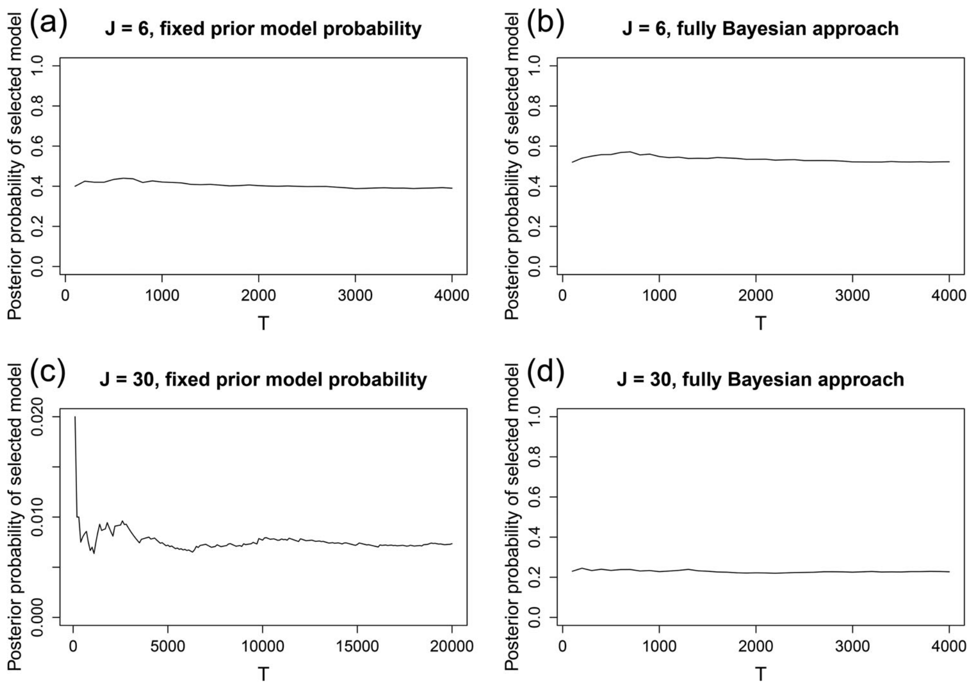

Figure 2. Convergence of the Gibbs sampler chain.

Table 2. Comparison of exhaustive calculation and MCMC model search method under one-sided variable selection.

	Fixed		Fully Bayesian		Fixed		Fully Bayesian	
	$g=n$		$g=n$		$g=J^2$		$g=J^2$	
J	Exh.	MCMC	Exh.	MCMC	Exh.	MCMC	Exh.	MCMC
6	0.394	0.404	0.523	0.531	0.425	0.430	0.502	0.509
8	0.450	0.452	0.434	0.432	0.247	0.248	0.442	0.439
10	0.436	0.433	0.712	0.718	0.437	0.433	0.712	0.718
12	0.289	0.292	0.665	0.676	0.346	0.346	0.715	0.722
14	0.182	0.178	0.622	0.628	0.268	0.260	0.707	0.709

"Fixed" denotes fixed prior model probabilities. "Fully Bayesian" denotes fully Bayesian approach. "Exh." denotes exhaustive calculation. "MCMC" denotes MCMC model search method.

of 0.5, 0.3 and 0, respectively, in the population above.

The model search method using the MCMC samples as introduced in Section 5 will be used to obtain a sample of γ' from which the best model can be determined. First, we have to discard the burn-in phase and check the convergence of the chain. It is not advisable to monitor the sample of γ' because it is a vector of discrete variables which necessarily fluctuates in the chain. Instead, we monitor the largest posterior probability among all possible models given the current sample, since it is the criteria to select the best model. The Gibbs sampler chain is checked per 100 samples. For example, if in the first 100 samples $M_{\gamma'}$ appears most often, say 40 times, then the probability is 0.4. Sequentially, if for the first 200 samples $M'_{\gamma'}$ (which is often the same as $M_{\gamma'}$) shares the largest count, say 100, then the probability becomes 0.5. As the number of iterations in the Gibbs sampler increases, the largest posterior probability should converge to a certain value such that we can safely select the best model. This can be best verified graphically. Figure 2 depicts the posterior probability of the selected model against the iteration number for γ' given $J=6$ and $J=30$ under $g=n$ using both the fixed setting and the fully Bayesian approach for the prior model probabilities. The chain starts with the null model $\gamma' = \boldsymbol{0}$. As can be seen from Figure 2 (a,b) with $J=6$ candidate variables, the chain converges fast, that is, with more than 2,000 iterations the posterior probabilities become stable. While with $J=30$ variables, the chain would need more iterations, say, 10,000, to converge under the fixed prior probabilities, which is shown in Figure 2(c). However, when using the fully Bayesian approach, the chain needs much less iterations to converge even with $J=30$, which can be seen from Figure 2(d), because the posterior probability of the selected model obtained from the fully Bayesian approach is much higher than that from the fixed prior probabilities in the one-sided variable selection.

After checking the convergence of the Gibbs sampler, we discard 10,000 burn-in iterations and sample another 10,000 iterations for variable selection. To investigate the performance of the model search method, we compare it to exhaustive calculation (i.e., consider every possible set of variables) in terms of the posterior probabilities of the true model given a random sampled dataset. Table 2 presents the

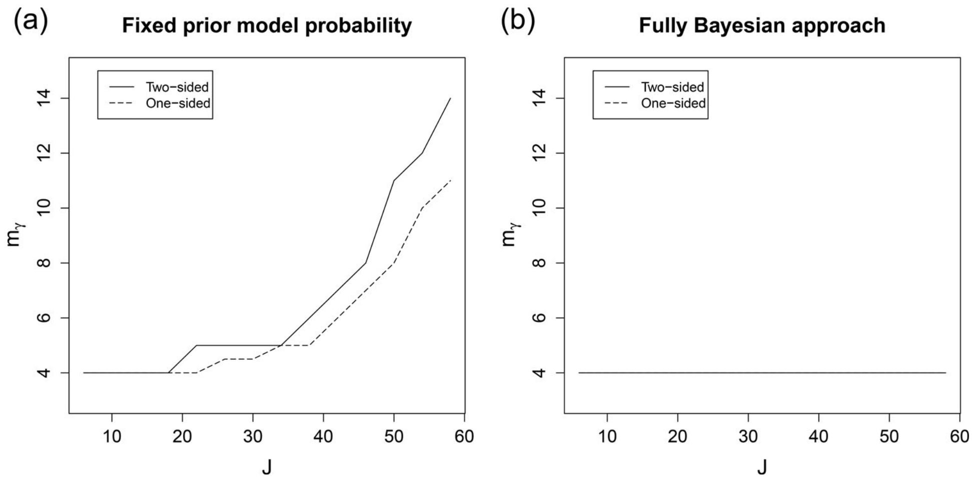

Figure 3. Variable selection based on fixed prior model probabilities and prior model probabilities specified using the fully Bayesian approach.

posterior probability of the true model that includes the first four variables under the fixed and fully Bayesian specifications of prior probabilities given $g = n$ and $g = J^2$ for $J = 6, 8, 10, 12, 14$. Note that exhaustive calculation is infeasible for a larger J. As can be seen, the posterior probabilities obtained from exhaustive and MCMC model search methods are very similar. This implies that the model search method performs very well for one-sided variable selection. In addition, it is interesting to note that with the increase of J the posterior probability of the true model decreases in general under the fixed prior probability because more models are under consideration. However, the fully Bayesian approach still results in quite large posterior probabilities for the true model.

Next, we illustrate the multiplicity problem given a large number of variables, and its correction using the fully Bayesian approach. Based on the sampling scenario (31), all variables except the first four are irrelevant to the outcome variable. Therefore, we would expect that the number of included variables is consistently around $m_\gamma = 4$ across different J. To reduce the sampling error, we simulate the data x_{ji} and y_i 100 times and report the median of the number of included variable. Note that the median is a more appropriate statistic than the mean because it will not be affected much by extreme values.

The number of selected variables m_γ against the number of candidate variables J based on fixed prior model probabilities is plotted in Figure 3(a), where J is varied by $J = 8, 12, ..., 56, 60$. As can be seen, with the increase of J the number of selected variables has an apparent increase when $J > 20$ under both selection approaches. This means that variable selection will select more variables regardless of the actual number of effective variables. For example, given $J = 60$ the two-sided and one-sided variable selection approaches result in the inclusions of 14 and 11 variables on average, respectively, which are much more than the number of true variables $m_\gamma = 4$ in the simulation. Therefore, the multiplicity problem happens when comparing a great many models.

As was presented in Section 4.2, multiplicity can be controlled using the fully Bayesian approach for the prior model probability specification. The number of included variables against the number of candidate variables based on the fully Bayesian approach is plotted in Figure 3(b). From this figure, we can clearly observe that both the two-sided and one-sided variable selection approaches render $m_\gamma = 4$ variables all the time. Therefore, it can be concluded that the fully Bayesian approach performs very well in terms of controlling multiplicity when the number of candidate variables is large.

General recommendations

Based on the simulation studies, four recommendations are given for Bayesian variable selection.

1. Bayesian one-sided variable selection is generally recommended.
2. The use of informative priors for the direction of the effects is recommended. The degree of prior information can be precisely tuned based on the amount of prior certainty.

Table 3. Summary of variables and coefficients.

Variables	Mean	SD	Coefficient	SE	p-value
0. Reading test score	655.0	20.11			
1. Total enrollment	2628.8	3913.1	0.002	0.003	0.374
2. Number of teachers	129.1	187.9	−0.075	0.058	0.200
3. Number of computers	303.4	441.3	0.003	0.005	0.520
4. Computer per student	0.136	0.065	32.29	12.40	0.010
5. Expenditure per student	5312	633.9	−0.003	0.001	0.045
6. Student teacher ratio	19.64	1.89	−0.803	0.527	0.128
7. District average income*	15.32	7.23	1.943	0.099	0.000

*Income in thousands of dollars.

3. The MCMC model search method is recommended if the number of candidate variables is larger than 15.
4. Prior model probability specification based on the fully Bayesian approach is recommended if the number of candidate variables is large to correct for multiplicity in a natural manner.

The proposed Bayesian one-sided variable selection approach will be demonstrated by two empirical data examples introduced in the next section.

Empirical data examples

In this section, two real data examples are used to illustrate Bayesian one-sided variable selection. The first one demonstrates the use of the one-sided approach under different specifications of the prior model probabilities. The second focuses on the use of the MCMC model search method and the fully Bayesian approach for a relatively large number of candidate variables.

Example 1

The first example concerns a dataset previously used by McNeish (2015) which contains average reading test scores from 420 schools (K6 and K8) in California for the 1998–1999 school year. This dataset is available in the R package Ecdat and called ”Caschool”. Besides the test score the dataset also includes a number of variables such as geographical information (e.g., county and district) and school characteristics. In this example, we consider seven school characteristics as possible predictor variables. Table 3 displays the outcome and predictor variables with means and standard deviations. In addition, the last three columns of Table 3 show the OLS estimates, standard errors, and p-values of the corresponding coefficients for the seven candidate variables.

We consider three different prior beliefs of the effect directions of the seven variables:

Table 4. Bayesian variable selection results for Example 1.

	Fixed prior probability		Fully Bayesian approach	
$n = 420$	Variables[a]	PMP[b]	Variables	PMP
Two-sided	{2,4,7}	0.349	{2,4,7}	0.323
One-sided (i)	{−1,4,7}	0.788	{−1,4,7}	0.709
One-sided (ii)	{−1,4,7}	0.347	{−1,4,7}	0.572
One-sided (iii)	{−2,4,7}	0.341	{−2,4,7}	0.252

[a] Variables included: minus means negative effect.
[b] PMP: posterior model probabilities.

i. prior belief $= (n\pm, s+, s+, s+, s+, s-, n\pm)$
ii. prior belief $= (n\pm, m+, m+, m+, m+, m-, n\pm)$
iii. prior belief $= (n\pm, n\pm, n\pm, n\pm, n\pm, n\pm, n\pm)$

where (i) and (ii) suggest strong and moderate prior beliefs, respectively, of the effect directions for some variables, and (iii) suggests no prior belief for all variables. The direction of the effect for each variable is determined by our prior knowledge. In this example, variables ”Number of teachers”, ”Number of Computers”, ”Computer per student”, and ”Expenditure per student” would have positive effects, whereas variable ”Student teacher ratio” is expected to have a negative effect on students' average test score (based on previous studies of the relationship between student academic performance and school expenditure). When we have no idea about the direction of a variable, e.g., ”Total enrollment” and ”District average income”, then ”$n\pm$” is given. Note that the directions in the prior beliefs should be specified without referring to the signs of the OLS estimates. Thereafter, we will select a subset of variables that best predicts the reading test score using the Bayesian one-sided variable selection with the fixed and fully Bayesian specifications of prior model probabilities.

The selection outcome based on the two-sided and one-sided approaches with prior beliefs (i), (ii) and (iii) are shown in Table 4 given $g = n$. As can be seen, the one-sided approach with no prior belief (iii) results in the same included variables as the two-sided approach. Note however that the one-sided approach also provides the direction of the effects. With prior beliefs (i) and (ii) the one-sided approach selects different variables than the two-sided approach. Furthermore, with strong prior belief (i) the posterior probabilities of the selected model are larger than those under the moderate prior belief (ii). This means that the best model can be chosen with more certainty if strong prior beliefs are present about the direction. Finally, we can conclude that variables “Total enrollment”, “Computer per student” and “District average income” are selected when researchers have informative priors of the effect directions.

Table 5. Bayesian variable selection results for Example 2.

	Fixed prior probability Variables[a]	PMP[b]	Fully Bayesian approach Variables	PMP
Two-sided	{3,4,8,14,15,16}*	0.041	{3,4,8,15,16}	0.064
One-sided	{3,4,–8,14,–15,16}	0.022	{3,4,–8,–15,16}	0.007

[a] Variable included: minus means negative effect.

[b] PMP: posterior model probabilities.

*3. Net assets; 4. Academic numbers; 8. Technician numbers; 14. Furniture and equipment; 15. Land and buildings; 16. Research grants.

Example 2

The second example comes from a study of efficiency in the provision of $n = 62$ UK universities. The data is again available in the R package Ecdat and named ”University”. The outcome variable is the university research rank with $J = 16$ candidate predictor variables including, e.g., academic numbers, academic pay, and land and buildings. Our target is to select a subset of variables to estimate the efficiency of the cost in the 62 UK universities, which can be achieved by Bayesian one-sided variable selection based on the MCMC model search method.

Both fixed prior probabilities and prior probabilities specified using the fully Bayesian approach are used. In this example, we assume no prior belief available for the effect directions. The Bayesian variable selection results are shown in Table 5. As can be seen the two-sided and one-sided variable selections lead to the same set of variables included, because no prior belief of the direction is involved in the one-sided approach. More interestingly, the fully Bayesian approach results in fewer included variables than the fixed setting, because the multiplicity is controlled. Finally, we recommend including variables (3, 4, 8, 15, 16) in the regression model, where the third, fourth and sixteenth variables have positive effects and the eighth and fifteenth have negative effects.

Conclusion

This paper proposed a Bayesian one-sided variable selection scheme which can incorporate the prior beliefs of the effect directions of predictor variables. When the number of candidate variables is large, the MCMC model search method which only visits models with high posterior probabilities has been adopted for fast computation. In addition, the fully Bayesian approach is useful for multiplicity correction in the one-sided selection.

From the simulation studies, several conclusions can be drawn. First, the one-sided Bayesian variable selection with correctly specified priors of the directions increases the chance of selecting the true model compared to the traditional two-sided approach in various cases, especially when the sample size and the effect sizes are relatively small. Second, the MCMC model search method performs quite well in terms of fast convergence and accurate selection. Third, the specification of prior model probabilities using the fully Bayesian approach can effectively control for multiplicity. Because our simulation studies only considered specific conditions and was not exhaustive (to keep the scope of the paper reasonable), we can only present general guidelines and recommendations as done at the end of Section 6. Nevertheless, Bayesian one-sided variable selection with the MCMC model search method and together with the fully Bayesian approach for prior model probability specification can offer researchers in the social and behavioral sciences a feasible, reasonable, and powerful technique in exploratory regression analysis.

The number of candidate models under one-sided variable selection is often larger than under two-sided variable selection, which results in a model splitting effect where models that are clearly incorrect receive some prior probability causing more posterior uncertainty. Depending on the specification of the prior probabilities, this could make the MCMC model search algorithm less effective. However, the method proposed in this paper can completely avoid or considerably reduce the model splitting effect. First, the proposed method allows researchers to specify zero prior probability to a negative effect of a variable x_j, if they believe that the effect must be positive if it is nonzero. Models with negative effect of x_j will have a prior probability of zero and will not be involved in the variable selection process. This would avoid the model splitting effect caused by variable x_j. Second, the prior inclusion probabilities under the two-sided and the one-sided variable selection approaches are equal. Given the same data, the posterior probability of including a variable under both approaches is therefore identical. Thus, there is no model splitting effect in terms of selecting a subset of variables to have a nonzero effect. Third, the fully Bayesian approach favors models with fewer included variables when the number of candidate variables is large. This further implies that the approach prefers to only include variables that have large effects. For these variables, there is little chance to obtain wrong effect directions. Thus, the promising models under a one-sided approach will have a similar amount of posterior probability mass as under a two-sided approach. In addition, the MCMC model search algorithm is still quite effective when using the fully Bayesian approach given a large number of candidate variables, which can be seen from Figure 2.

In this paper, the g prior specified for the regression coefficients is a local prior distribution with a mean of zero. We used the truncated g prior in the one-sided test because it implies that small effects (either positive or negative) are more likely than large effects, which is generally observed in the social and behavioral sciences. However, the non-local prior proposed by Johnson and Rossell (2010) would be another way for prior specification in the one-sided variable selection. Finally, this paper does not discuss the case that the sample size is less than the number of candidate variables, which itself is a challenging topic in Bayesian model selection. This would be an interesting setting to explore in further research.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant NWO 024.001.003 from the Netherlands Organization for Scientific Research, Grant NWO Vidi 452-17-006 from the Netherlands Organization for Scientific Research, and Grant 2019ECNU-XFZH015 from the East China Normal University.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Eric-Jan Wagenmakers and an anonymous reviewer for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the NWO is not intended and should not be inferred.

ORCID

Herbert Hoijtink http://orcid.org/0000-0001-8509-1973

References

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*(3), 870–897. https://doi.org/10.1214/009053604000000238

Berger, J., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94*(446), 542–554. https://doi.org/10.1080/01621459.1999.10474149

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*(2), 627–679. https://doi.org/10.1214/18-BA1103

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, *12*(1), 27–36. https://doi.org/10.1023/A:1013164120801

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, *22*(4), 1947–1975. https://doi.org/10.1214/aos/1176325766

Garcia-Donato, G. & Martinez-Beneito, (2013). On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, *108*, 340–352.

George, E. I., & Foster, D. P. (2000). Calibration and empirical Bayesian variable selection. *Biometrika*, *87*(4), 731–747. https://doi.org/10.1093/biomet/87.4.731

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*, 339–373. https://www.jstor.org/stable/24306083

Gu, X., Mulder, J., & Hoijtink, H. (2016). Error probabilities in default bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130–143. https://doi.org/10.1016/j.jmp.2015.09.001

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fraction Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Hughes, A., & King, M. (2003). Model selection using AIC in the presence one-sided information. *Journal of Statistical Planning and Inference*, *115*(2), 397–411. https://doi.org/10.1016/S0378-3758(02)00159-3

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Johnson, V., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 143–170. https://doi.org/10.1111/j.1467-9868.2009.00730.x

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kass, R., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, *90*(431), 928–934. https://doi.org/10.1080/01621459.1995.10476592

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Kuo, L., & Mallick, B. V. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, *60*, 65–81. https://www.jstor.org/stable/25053023

Liang, F., Paulo, R., Molina, G., Clyde, M., & Berger, J. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. https://doi.org/10.1198/016214507000001337

Marsman, M., & Wagenmakers, E. (2017). Three insights from a Bayesian interpretation of the one-sided P value. *Educational and Psychological Measurement*, *77*(3), 529–539. https://doi.org/10.1177/0013164416669201

McNeish, D. M. (2015). Using Lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–483. https://doi.org/10.1080/00273171.2015.1036965

Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *The British Journal of Mathematical and Statistical Psychology*, *67*(1), 153–171. https://doi.org/10.1111/bmsp.12013

Nott, D., & Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika*, *92*(4), 747–763. https://doi.org/10.1093/biomet/92.4.747

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Ročková, V., & George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, *109*(506), 828–846. https://doi.org/10.1080/01621459.2013.869223

Ročková, V., & George, E. I. (2018). The spike-and-slab Lasso. *Journal of the American Statistical Association*, *113*(521), 431–444. https://doi.org/10.1080/01621459.2016.1260469

Scott, J., & Berger, J. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*(5), 2587–2619. https://doi.org/10.1214/10-AOS792

Stock, J. H., & Watson, M. W. (2015). *Introduction to econometrics*. Addison-Wesley.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600–620. https://doi.org/10.1080/01621459.2015.1108848

van Erp, S., Oberski, D., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Wolak, F. A. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association*, *82*(399), 782–793. https://doi.org/10.1080/01621459.1987.10478499

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