Bayesian Variable Selection for Linear and Nonlinear Effects in Regression Models

Afra Kilic

Supervisor: dr. ir. Joris Mulder

Abstract

A computationally efficient Bayesian variable selection method is presented for linear and nonlinear effects in regression models using the Laplace approximation of the marginal likelihood via Bayesian Information Criteria (BIC). The proposed method can be used to evaluate the inclusion of variables as either linear or nonlinear predictors in the model, as well as to identify redundant variables for removal. The best subset of variables is selected by comparing the posterior probabilities of the possible models. In case of large number number of candidate variables, a Bayesian model search method that only visits models with high posterior probabilities is adopted for fast computation, and multiplicity is controlled by setting a Dirichlet prior on the probabilities of observing each effect type. The performance of the proposed method is investigated using several simulation studies and two real data examples. The R code used in this study and for applying the proposed method is provided.

1. Introduction

Appropriate variable selection is a fundamental step when analyzing data collected for social and behavioral studies. In regression models, the objective of variable selection is to identify the relevant predictor variables of the outcome variable and remove the redundant ones before conducting the analyses as they will add noise when estimating or testing other quantities of interest. Variable selection is traditionally a problem of determining the relevance of predictor variables by assessing whether their coefficients significantly deviate from zero in relation to the outcome variable. To do so, most of the existing selection methods assume linearity between explanatory and dependent variables. However, real life data can often contain a mix of both linear and nonlinear predictor variables, along with redundant variables. In other words, a predictor variable can have a non-zero linear effect, a non-zero nonlinear effect, or no effect at all on the outcome variable.

The literature on variable selection methods for general linear models is too vast to summarize. The most popular modern methods include the least absolute shrinkage and selection operator (Lasso), introduced by Tibshirani (1996) in the frequentist setting, variable selection using "spike and slab" priors (George and McCulloch, 1993) and their mixture Bayesian Lasso has been developed by Park and Casella (2008) and Rockova and George (2018). For nonlinear models, variable selection becomes challenging due to the absence of closed-form solutions for maximum likelihood estimates. However, recent years have seen progress in developing algorithms for nonlinear regression, including SpAM (Ravikumar et al., 2009), HSIC Lasso (Yamada et al., 2014), stepwise selection (Peduzzin, Hardy, and Holford, 1980), FVM (Li, Yang, and Xing, 2005), hierarchical multiple kernel learning (Bach, 2008), and RGS (Navot et al., 2005) in the frequentist setting. In the Bayesian setting, variable selection methods for logistic regression using power posterior (Maity, 2016) and for multivariate nonlinear regression with graph structures using knot splines (Niu et al., 2020) have been proposed.

However, the existing variable selection methods for linear or nonlinear models in both frequentist and Bayesian frameworks do not inform us about the true nature of the relationship between the variables. In practice, researchers often rely on visual

inspection, such as scatter plots, to assess the relationship between variables. When a possible nonlinear relationship is observed, linear transformations like polynomials or logarithmic functions are applied to those variables to create approximate linearity. This procedure is problematic for several reasons. First, performing several different significance tests on different transformed variables may lead to inflated Type-I error and p-hacking. Second, analyzing the linear relationship between transformed (X^2) and the outcome (Y) variables fails to inform us about the direct relationship between the variables. Third, transformations are only able to create approximate linearity for a limited set of nonlinear relationships. Fourth, even if appropriate nonlinear models are used instead of linear transformations, identifying/understanding the form of the nonlinear relationship can be a very difficult task. Finally, visual inspection can be subjective to decide the type of the true relationship, and thus a principled and probabilistic approach is needed.

To address this limitation, a Bayes factor to test the nature of the relationship between the predictor variable and the outcome variable (linear vs nonlinear) using Gaussian Process prior with a square kernel has been proposed by Mulder (2022). Compared to conventional significance tests via p-value, this method can be used to quantify the statistical evidence in favor of linearity. However, it is only applicable for testing the effect of a single predictor variable and requires substantial computation time. In this study, we extend the Bayesian approach for testing linear vs. nonlinear relationships to the variable selection problem in regression models by using the Laplace approximation of the marginal likelihood via Bayesian Information Criterion (BIC). The proposed method can be used to evaluate the inclusion of variables as either linear or nonlinear predictors during the construction of statistical models, as well as determining their redundancy and subsequent removal from the model.

Under the Bayesian approach, the formal way to select the best subset of variables along with their effect types is to select the model with the highest posterior probability. Using this fact the problem may be thought as an optimization problem over the model space where the objective function is the posterior probability of the model and the maximization is taken place with respect to the models. The posterior model probability is proportional to the marginal likelihood of the data for the model of interest times the prior model probability. Laplace approximation of the marginal

likelihood via BIC is used to estimate the posterior model probabilities with fixed prior model probabilities. Furthermore, as it becomes computationally infeasible to evaluate all the models in the model space, we extend the Markov chain Monte Carlo (MCMC) model search method proposed by George and McCulloch (1997) to the nonlinear variable variable selection problem. In practice, most candidate models have almost zero posterior probability and will not be visited by the algorithm. Hence, the algorithm will visit only those models with relatively high posterior probabilities, reducing computation time substantially.

The paper is organized as follows: In Section 2,the variable selection scheme and generalized additive models are described. Section 3 presents an explanation of how the Bayes factor is approximated using BIC and integrated into the variable selection process. Thereafter, MCMC model search algorithm is introduced for the larger number of variables in Section 4. The fifth section continues with a simulation to evaluate the consistency of the proposed Bayes factor and the performance of the MCMC model search algorithm. Subsequently, two empirical data examples are used to illustrate how the proposed variable selection scheme can be used in Section 6. The seventh section concludes.

2. Variable Selection Scheme

In this paper, we consider a variable selection problem in the context of generalized additive models (GAMs). GAMs are a type of generalized linear model that allow for flexible, non-linear relationships between the predictor variables and the response variable. GAMs accomplish this by modeling the response variable as a sum of smooth functions of the predictor variables, rather than as a linear combination. This allows for more complex and nuanced relationships to be captured, which can be particularly useful when the true relationship between variables is not well understood or linear.

A generalized additive model for a response variable Y, modeled as a function of predictor variables $x_1, x_2, ..., x_j$ can be expressed as:

$$g(\mu_i) = X_i\beta^* + f_1(x_{1i}) + f_2(x_{2i}) + ... + f_j(x_{ji}) + \epsilon \tag{1}$$

where $g(.)$ is the link function, $\mu_i = E(Y_i)$, $X_i\beta^*$ is a row of the model matrix

for any strictly parametric model components, β is the corresponding parameter vector. f_i are smooth functions of the predictor variables x_i, j is the number of variables and $\epsilon \sim N(0, \sigma^2)$ are the residuals with σ^2 being the variance. The function, f_i, can be modeled using a variety of smoothing methods, such as splines, kernel methods, or generalized cross-validation. Notice that with no smooth function, $g(.)$ can be considered as a linear model. Our target is to identify the nonlinear and linear components of $g(\mu_i)$ and remove the redundant covariates by testing all possible models in the model space against each other. For instance, consider a two-variable model $y = f_1(x_1) + \beta_2 x_2 + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. f_1 defines the nonlinear relationship between x_1 and y; $\beta_2 = a \in \mathbb{R}$ stands for the (non)zero (linear) relationship between x_2 and y. There are two predictor variables ($j = 2$) that can have three possible different effect on the outcome variables. Thus, 3^2 possible models need to be specified;

$$
\begin{array}{lll}
M_0 : Y \sim 1 & M_1 : Y \sim x_1 & M_2 : Y \sim s(x_1) \\
M_3 : Y \sim x_2 & M_4 : Y \sim s(x_2) & M_5 : Y \sim x_1 + x_2 \\
M_6 : Y \sim s(x_1) + x_2 & M_7 : Y \sim x_1 + s(x_2) & M_8 : Y \sim s(x_1) + s(x_2)
\end{array} \tag{2}
$$

where the smooth term, $s(.)$, captures the nonlinear relationship between the predictor and outcome variables within generalized additive models. For instance, model M_6 indicates the presence of both a nonzero nonlinear effect of x_1 and a nonzero linear effect of x_2 on the outcome variable. We expect the true model to have the largest posterior probability among the model space. A straightforward advantage of this approach is that besides selecting a set of variables that best predicts the outcome variable, the effect type of the 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 a linear or nonlinear effect on the outcome variable and with how much certainty determined by the posterior probability of the selected model.

In this study, thin-plate regression spline (Wood, 2003), the default choice implemented in the `gam` function of the R package `mgcv` (Wood, 2023), is used as the smooth term. Thin plate regression spline estimates a smoothing function f that minimizes a penalized least squares function g:

$$g(x_j, y, \lambda_j) = \sum_{i=1}^{N} (y_i - f(x_{ji})^2) + \lambda_j P(f) \tag{3}$$

where x_j is the predictor variable, y is the outcome variable, N is the sample size and P is a function that penalizes how wiggly/complex the function f is. λ_j is the penalization parameter and estimated via generalized cross validation (GCV); As $\lambda \to \infty$, the result is a linear fit and as $\lambda \to 0$ we have the opposite effect where any wiggliness is incorporated into the model. To use this function we also need to specify `k` which controls the number of basis function for f and thus is another way to control the degree of wiggliness of the function. The exact choice of `k` is not generally critical; it should be chosen to be sufficiently large to capture the underlying nonlinear relationship, yet small enough to maintain acceptable computational efficiency. We specify the smooth function with `k=4` to ease the computation and it is sufficient to capture the nonlinear relationships defined in this study.

3. Bayesian Variable Selection

A common strategy in Bayesian variable selection is involving an indicator variable (George & McCulloch, 1993), often denoted by $\gamma = (\gamma_1, ..., \gamma_J)$, where $\gamma_j = 0$ implies zero effect of x_{ji}; $\gamma_j = 1$ implies a non-zero linear effect of x_{ji}, and $\gamma_j = 2$ implies a non-zero nonlinear effect of x_{ji} on the outcome variable. Since each λ_j has three possible values, there are 3^J candidate models each of which can be denoted by M_λ.

The Bayes factor is a criterion for comparing two hypotheses or models and can be used to select the true model in the model space. For example, consider the model space in (2), we consider the null model M_0 as the base model and compute Bayes factors of candidate models against the base. In this case Bayes factors are the ratio of marginal likelihoods of the data under candidate models and and the null model (Kass & Raftery, 1995):

$$BF_{\gamma 0} = m(Y|M_\gamma)/m(Y|M_0). \tag{4}$$

where $m(Y|M) = \int (m(Y|M, \theta)\pi(\theta|M)d\theta$ and Y denotes the observed data and θ denotes the parameter space, $m(Y|M, \theta)$ is the likelihood for model M and $\pi(\theta|M)$ is

the prior density of θ. Bayes factor is a summary evidence for model M_γ against the null model M_0; higher values of $BF_{\gamma 0}$ refers to more evidence for M_γ whereas $BF_{\gamma 0}$ values closer to zero refers to higher evidence for M_0. For instance, $BF_{\gamma 0} = 5$ means that it is five times more likely to to have observed the data under M_γ than the null model M_0. As Bayes factors can span a wide range of values (i.e., $BF \in (0, \infty)$), $log(BF_{\gamma 0})$ will be considered to interpret and compare the evidence for each model in a more standardized manner. For more information about Bayes factor interpretation, see the evidence categories suggested by Jeffreys (1961).

In order to compute Bayes factor, priors have to be specified for the unknown model parameters for each model before observing the data. For testing problem considered in this work, defining the parameter space for the GAMs is not straightforward as the number of free parameters in GAMs with smoothing splines are controlled by the degree of penalization selected during fitting, by GCV. Furthermore, given the complexity of GAMs, computation of the marginal likelihood is a challenging task. Hence, the LaPlace approximation of Bayes factor via BIC will be considered in this study. This method has computational advantage and it does not require to specify prior distribution because it implicitly assumes unit information prior (Wagenmakers, 2007). The unit information prior was proposed by Kass and Wasserman (1995) based on Fisher information and is a multivariate normal prior with a mean at the maximum likelihood estimate and variance equal to the expected information matrix for one information (Kass and Wasserman, 1995). The unit information prior is well spread out compared to the likelihood and is relatively flat within the parameter space where the likelihood is relevant, without being much larger outside of that region. Hence, the likelihood dominates the prior and it satisfies the conditions for a *stable estimation situation* where resulting inference is relatively insensitive to the prior (Edwards, Lindman and Savage, 1963). This property ensures that the BIC is "objective" in the sense that different researchers will have the same statistical inference from the same data and set of models. Typically, BIC for M_γ can be expressed as

$$BIC(M_\gamma) = -2logL_\gamma + df_\gamma log(N) \tag{5}$$

where N is the number of observations, df_γ is the number of free parameters of model M_γ, L_γ is the maximum likelihood for model M_γ, that is $L_\gamma = m(Y|M_\gamma, \theta)$.

The BIC can be used approximate the marginal likelihood as

$$m(Y|M_\gamma) = exp[-BIC(M_\gamma/2)] \tag{6}$$

In case of two models or hypotheses, the Bayes factor is defined as the ratio of the prior predictive probabilities as

$$BF_{\gamma 0} \approx \frac{m(Y|M_\gamma)}{m(Y|M_0)} = exp(\triangle BIC_{\gamma 0}/2) \tag{7}$$

where $\triangle BIC_{\gamma 0} = BIC(M_0) - BIC(M_\gamma)$. For the derivation of BIC approximation of Bayes factor and its details we refer to Raftery(1995) and Wagenmakers (2007).

As mentioned above, calculating model degrees of freedom, df_i, is not straightforward for GAMs with a smoothing spline. In practice `k-1` sets the upper limit on the degrees of freedom associated with an `s()` smooth (1 degree of freedom is usually lost to the identifiability constraint on the smooth). However the actual effective degrees of freedom are controlled by the degree of penalization selected during fitting, by GCV. In the case of the default smooth term with `k=4`, `s(x,k=4)`, 3 basis functions are used to estimate the effect of x on the outcome variable. The upper limit for effective degrees of freedom (edf) can be at most 3 based on the penalty and edf=1 represents a linear relationship between the variables.

`logLik` function built under the R package "stats" is used to extract L_γ for each model and it is applicable to both `lm` and `gam` objects. If a `gam` object includes smooth term, `logLik` corrects the degrees of freedom based on the effective degrees of freedom. When testing a linear relationship with an `s()` smooth, `gam` is expected to result in a linear fit with the same L_γ and df_γ as in `gam` without the smooth term. In other words, when the true relationship between x and y is linear, BIC for the models $Y \sim x$ and $Y \sim s(x)$ will be the same meaning that Bayes factor between them will be 1. However, `gam` with the smooth term is still a more complex model because compared to a linear model fit we are estimating 2 more free parameters in the case of `k=4`. The principle of Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. For the variable selection problem considered in this work, the model with the smoothing spline is needed to be penalized to choose

the simpler model. Thus, edf is fixed to its maximum value 3 for each smooth which is supposed to be 1 for a true linear relationship.

Using the Bayes factor obtained in (7) the posterior probability of each model can be computed as

$$P(M_{\gamma'}|Y) \propto P(M_{\gamma'})BF_{\gamma'}0 \tag{8}$$

where $P(M_{\gamma'}) = \frac{1}{3^J}$ is fixed for each model. The model having the largest posterior probability will be selected.

4. MCMC Model Search Method

When the number of candidate variables is large, exhaustive calculation of the posterior model probabilities in (8) for all possible models becomes infeasible. For instance, given $J = 20$, there are more than three billions (3^{20}) possible models under consideration. This problem can be addressed by using MCMC model search method in the selection algorithm with different prior model probability setting.

A popular MCMC model search method was proposed by George and McCulloch (1993) for cases where the model space is large. The basic idea of the MCMC algorithm for Bayesian variable selection is to sequentially sample γ from its posterior distribution, $\pi(\gamma|Y)$, and select the best model which appears most often in the sample of γ. The marginal posterior distribution of γ has an analytical form:

$$\pi(\gamma|Y) = P(M_\gamma|Y) \propto BF_{\gamma 0}P(M_\gamma) \tag{9}$$

where $BF_{\gamma 0}$ is given in (7) and $P(M_\gamma) = \frac{1}{3^J}$.

Gibbs sampler algorithm is only applied to γ, i.e.i to sequentially sample along γ_j^t for $j = 1, ..., J$ and $t = 1, ..., T$ with T the iteration number:

$$\gamma_1^0, ..., \gamma_J^0, \gamma_1^1, ..., \gamma_J^1, ..., \gamma_1^t, ..., \gamma_J^t, ..., \tag{10}$$

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

The conditional distribution of γ_j given all other γs is Bernoulli (George & McCulloch, 1997). The three probabilities of sampling $\gamma_j^t = r$ for $r = 0, 1, 2$ at iteration rate t are

$$P(\gamma_j^t = r|\gamma_{-j}^t, y) = \frac{\pi(\gamma_j^t = r|\gamma_{-j}^t, y)}{\sum_r \pi(\gamma_j^t = r|\gamma_{-j}^t, y)} \tag{11}$$

where $\gamma_{-j}^t = (\gamma_1^t, ..., \gamma_{j-1}^t, \gamma_{j+1}^{t-1}, ..., \gamma_J^{t-1})$ denotes the latest values of γ except γ_j. Note that when sampling $\gamma_j^t, (\gamma_{t+1} + ... + \gamma_J)$ have not been sampled at iteration t, and thus their values at the $t-1$ iteration are used. $\pi(\gamma_j^t = r|\gamma_{-j}^t, y)$ can be computed using Equation (9). However, regardless of the number of variables J, a fixed $P(M_\gamma)$ causes the algorithm to include more variables as J increases. This phenomenon, called multiplicity, arises from multiple tests or comparison in variable selection. This is most obvious in orthogonal situation where J independent tests on the effect type of the variable x_j are performed. Therefore, for instance, at iteration rate t, three independent models are compared to test the effect type of variable x_j on the outcome variable, and the prior model probabilities for each model will be $P(M_{\gamma_j}^t) = 1/3$. Regardless of J, $P(M_{\gamma_j}^t)$ remains the same at every iteration for each variable. However, $P(M_{\gamma_j}^t) = 1/3$ suggests a model size of $2J/3$ a priori since each variable has a probability of $2/3$ of being included (either as linear or nonlinear), and there are J total number of variables. This problem remains in case of other fixed prior choices, therefore no fixed choice of prior which is independent from the total number of variables can adjust for multiplicity.

To correct for the multiplicity, we specify a Dirichlet distribution, $Dirichlet(\alpha_0, \alpha_1, \alpha_2)$, for the prior probabilities of effect types. For variable x_j, prior probabilities of having nonlinear, linear and zero effects respectively are denoted by p_2, p_1 and p_0. $p_2 + p_1 + p_0 = 1$ and α_0, α_1, α_2 are the corresponding parameters for the Dirichlet distribution. Hence, Equation (12) can be rewritten by correcting multiplicity via multiplying each $\pi(\gamma_j^t = r|\gamma_{-j}^t, y)$ with p_r^t:

$$P(\gamma_j^t = r|\gamma_{-j}^t, p_r^t, Y) = \frac{\pi(\gamma_j^t = r|\gamma_{-j}^t, Y)p_r^t}{\sum_r(\pi(\gamma_j^t = r|\gamma_{-j}^t, Y)p_r^t)} \tag{12}$$

where p_r^t can be written as:

$$p_r^t \sim dirichlet(\alpha_0^{t-1} + |G_0^{t-1}|, \alpha_1^{t-1} + |G_1^{t-1}|, \alpha_2^{t-1} + |G_2^{t-1}|) \tag{13}$$

G_0, G_1, and G_2 are the numbers of variables of which the effect types are zero, linear and nonlinear respectively, and they are set to zero at $t = 0$. Note that $G_0 + G_1 + G_2 = J$ when $t \neq 0$. First, three sampling probabilities, either $\gamma_j^t = 0$, $\gamma_j^t = 1$ or $\gamma_j^t = 2$ will be sampled using Equation (12). Thereafter, the algorithm visits the next γ_{j+1}^t. Once all γ^t have been sampled, at the end of iteration t, the algorithm samples for p_r^{t+1} using the resulted γ^t. Then, the algorithm proceeds to the $t + 1$ iteration and sample for the next γ^{t+1} with updated prior probabilities, p_r^{t+1}, until the Gibbs chain converges to obtain the samples shown in (10). After obtaining the Gibbs samples and discarding the burn-in phase (e.g., the first 1000 iterations), the best model will be the one with the highest frequency in the useful samples.

Gibbs Sampler Algorithm: (*in total* $3 \times P \times T$ *model fit*)

itialize $\gamma^0 = 0$ and $p_r^0 = 1/3$ at $t = 0$

repeat

for $p = 1, ..., J$ do

Sample $\gamma^t = 0$ with probability $P(\gamma_j^t = 0|\gamma_{-j}^t, p_r^t, Y)$

Sample $\gamma^t = 1$ with probability $P(\gamma_j^t = 1|\gamma_{-j}^t, p_r^t, Y)$

Sample $\gamma^t = 2$ with probability $P(\gamma_j^t = 2|\gamma_{-j}^t, p_r^t, Y)$

end for

Sample p_r^{t+1} with probability $dirichlet(\alpha_0^t + |G_0^t|, \alpha_1^t + |G_1^t|, \alpha_2^t + |G_2^t|)$

set $t = t + 1$

until Gibbs chain converges.

5. Simulation Study

This section presents a two-stage simulation study to examine the consistency of the Bayes factor approximated via BIC and the performance of the MCMC model search method in case of a large number of candidate variables. In both scenarios, we assumed fixed model prior probabilities $\frac{1}{3^J}$.

I. Bayes Factor Consistency Check

In this part, we conducted a comprehensive simulation study to check whether BF approximated via BIC is consistent when identifying the effect types between the variables across various settings. By considering four different models as a null model (14.1), a linear non-zero model (14.2), a nonlinear non-zero model (14.3) and a comprehensive model consisting all types of relationships (14.4) the outcome variable y was generated as follow:

$$y \sim N(1, 0.1) \quad (1) \qquad y \sim \beta x_1 + \epsilon \quad (2) \qquad (14)$$
$$y \sim \beta f(x_1) + \epsilon \quad (3) \qquad y \sim \beta exp(x_1) + \beta x_2 + \beta_0 x_3 + \epsilon \quad (4)$$

where x_1, x_2 and x_3 are simulated independently from $N(1, .5)$; $\epsilon \sim N(0, \sigma^2)$ and $\beta_0 = 0$. $f(x_1)$ represents the nonlinear effect of x_1 on y and different nonlinear relationships such as sine, exponential, quadratic were specified to assess Bayes factor consistency across different nonlinear relationships. We investigated four different sample sizes: $n \in 50, 100, 500, 1000$. For the parameters β and σ^2, we generated two sequences of 10 values each, ranging from -3 to 3 for β and from 0.1 to 1 for σ^2. Each scenario was evaluated using 500 different datasets. The R code used for the simulation study can be found in **Appendix-A:**

First, the null model (14.1) is tested against a linear and a nonlinear models to check whether the Bayes factor can identify the zero effect across different sample sizes and σ^2. The following three models were tested against each other;

$$M_0 : Y \sim 1 \qquad M_1 : Y \sim 1 + x_1 \qquad M_2 : Y \sim 1 + s(x_1) \qquad (15)$$

where M_0 is the true model as there is no relationship between x_1 and y. Hence, we expect $log(BF_{01})$ and $log(BF_{02})$ to be greater than zero and result in favor of model M_0.

As the sample size n increases, both BF_{01} and BF_{02} also increase, indicating stronger evidence in favor of the null model (Fig. 1). On average, evidence against the nonlinear model is higher, which is expected as the nonlinear model has higher degrees of freedom with relatively the same loglikelihood value. The variability in the error term σ^2 does not systematically affect the relative evidence for the null

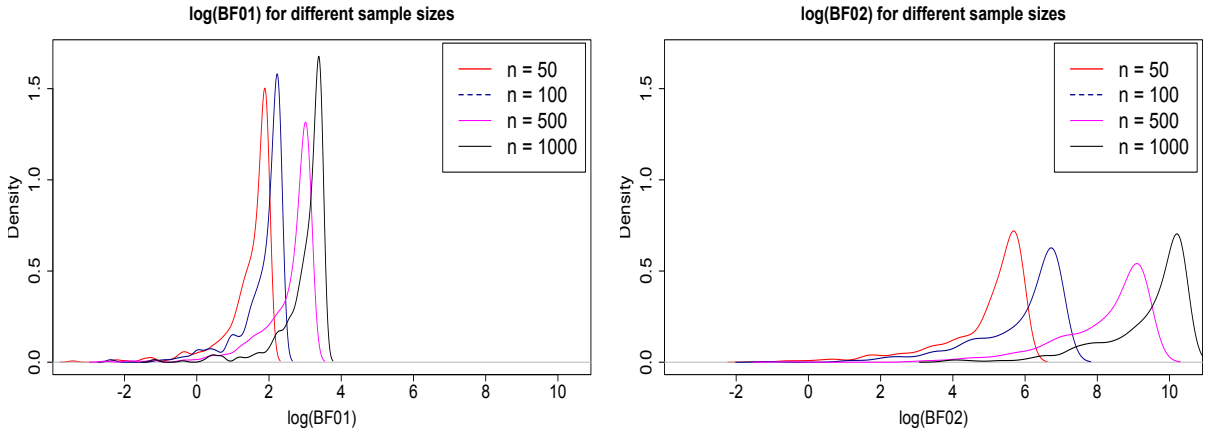

Figure 1: Logarithms of BF_{01}, B_{02} when sample size equals to 50, 100, 500 and 1000.

model against both the linear and nonlinear models.This observation aligns with expectations since there is no underlying relationship between x_1 and y. Therefore, changes in the variance of the outcome variable do not impact the results other than randomness (see

Second, the linear (14.2) and the nonlinear (14.3) models are tested to check whether the Bayes factor can identify the true linear and nonlinear effects across different sample sizes and effect sizes (β). Thus, for each model the following two models were tested against each other;

$$M_0 : Y \sim 1 + x_1 \qquad\qquad M_1 : Y \sim 1 + s(x_1) \tag{16}$$

For the linear model (14.2), $log(BF_{01})$ is expected to be greater than zero since the true relationship between x_1 and y is linear. Conversely, for the nonlinear model (14.3), $log(BF_{01})$ is expected to be lower than zero as the true relationship between x_1 and y is nonlinear. To represent various types of nonlinear relationships, we considered different functions for $f(x_1)$ in (14.3), including logarithmic, exponential, quadratic, sine, and a general nonlinear function. Specifically, we defined $f(x_1)$ as $log(x_1)$, $exp(x_1)$, $(x_1)^4$, $sin(\frac{x_1}{3})$, and $\phi(x_1)(x_1)$, where ϕ represents the standard normal probability density function. These choices allowed us to capture a diverse range of nonlinear relationships and assess the ability of the Bayes factor to differentiate between linear and nonlinear effects.

Results show that as n increases $log(BF_{01})$ increases when the true model is linear. Specifically, across 500 replications, average $log(BF_{01})$ increased from 3.36 when $n = 50$ to 6.40 when $n = 1000$ (see **Appendix-A**). When the true model is nonlinear, however, $log(BF_{01})$ decreases for all types of nonlinear relationships (Fig. 2a). The decrease in $log(BF_{01})$ when the true model is nonlinear is more drastic than the increase in $log(BF_{01})$ when the true model is linear. Thus, the Bayes factor can be considered better at identifying the nonlinear relationships.

When the true relationship is linear, β_1 does not have an effect on the Bayes factor. For the true linear relationships, nonlinear model, `gam` with the smoothing spline, will produce almost the same L_γ estimates in BIC (5) with `gam` without the spline. Consequently, the relative ratio of L_γ will remain unchanged across different values of β_1. However, as β_1 deviates from zero, the Bayes factor (BF_{01}) for the nonlinear models decreases, indicating an increasing evidence in favor of a nonlinear effect (Fig. 2b). Notice that even when $\beta_1 = 0$, the Bayes factor still favors the linear model. This is caused by the penalization applied to the nonlinear model by considering `df` =3 for each spline.

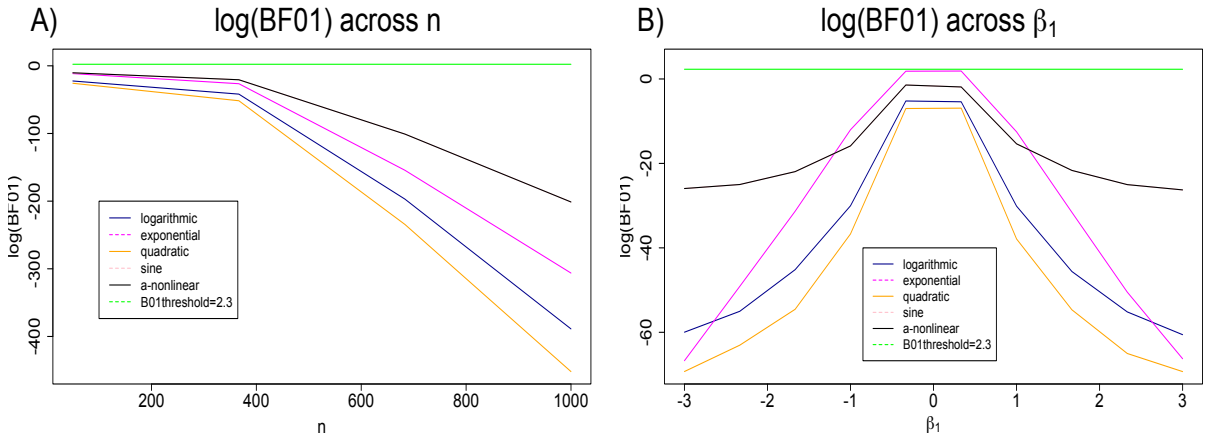

Figure 2: $log(BF_{01})$ **in y-axis for each type of nonlinear relationship across** $n \in \{50, 100, 500, 100\}$ **and** $\beta_1 \in seq(-3:3)$ **in x-axes.**

In the case of the model with three predictor variables (14.4), where x_1 has a non-zero nonlinear effect, x_2 has a non-zero linear effect, and x_3 has a zero effect on the outcome variable, a total of $3^3 = 27$ possible models were considered. The

true model, denoted as $M_T : Y \sim 1 + s(x_1) + x_2$, was expected to have the highest posterior probability among these models.

After conducting 500 replications, the algorithm successfully identified the true model, M_T, as having the highest posterior probability. Moreover, the average posterior probability for the true model increased as the sample size n increased and the effect sizes β_s moved further away from zero. However, when β_s approached zero, the algorithm struggled in capturing the nonlinearity in the data. In such cases, it tended to identify the nonlinear effects as linear effects, resulting in the true model being identified as $M_T : Y \sim 1 + x_1 + x_2$. Hence, while it can recognize the relevance of the variable x_1, it fails to identify its nonlinear effect on y. This suggests that as more data becomes available and the true effects become more pronounced, the evidence in favor of the true model becomes stronger.

II. Performance of the Model Search Method

In this subsection, we conducted a simulation study to assess the performance of the MCMC model search method. We considered five levels of number of variables $\{J \in 6, 10, 15, 20, 30\}$, and four levels of sample sizes $\{n \in 50, 100, 250, 500\}$, seven levels of effect sizes $\{\beta \in .1, .5, 1, 1.5, 2, 2.5, 3\}$ and five levels of number of knots $\{\texttt{k} \in 4, 6, 8, 10, 12\}$. For each scenario, 100 data sets were generated. The candidate variables $x_1, ..., x_J$ of length n were independently simulated from a standard normal distribution $N(0, 1)$. The nonlinear relationship is defined with the exponential function. Therefore, the outcome variable, y, was calculated with a zero intercept as follows:

$$y = \beta exp(x_1) + \beta exp(x_2) + \beta x_3 + \beta x_4 + \beta^0 X_{J-4} + \epsilon \tag{17}$$

where $\beta^0 = 0$, $\epsilon \sim N(0, \sigma^2)$ and $\sigma^2 = .1$. Thus, the first two variables exhibit a nonlinear effect, the next two variables have a linear effect, and the remaining variables have no effect on the outcome variable. The R code used for the simulation study can be found in **Appendix-B**.

The MCMC model search method as introduced in Section 4 will be used to obtain a sample of γ from which the best model can be determined. To start, we

need to discard the initial burn-in phase and ensure the convergence of the chain. Monitoring the sample of γ, which is a vector of discrete variables that fluctuates in the chain, is not recommended. Instead, we monitor the largest posterior probability among all possible models given the current sample since it serves as the criterion for selecting the best model. To check the Gibbs sampler chain, we examine every 100 samples. For instance, if in the first 100 samples, M_γ appears most frequently, say 40 times, then the probability is 0.4. Subsequently, if for the first 200 samples, $M_\gamma^{'}$ (which is often the same as M_γ) has 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, allowing us to confidently select the best model.

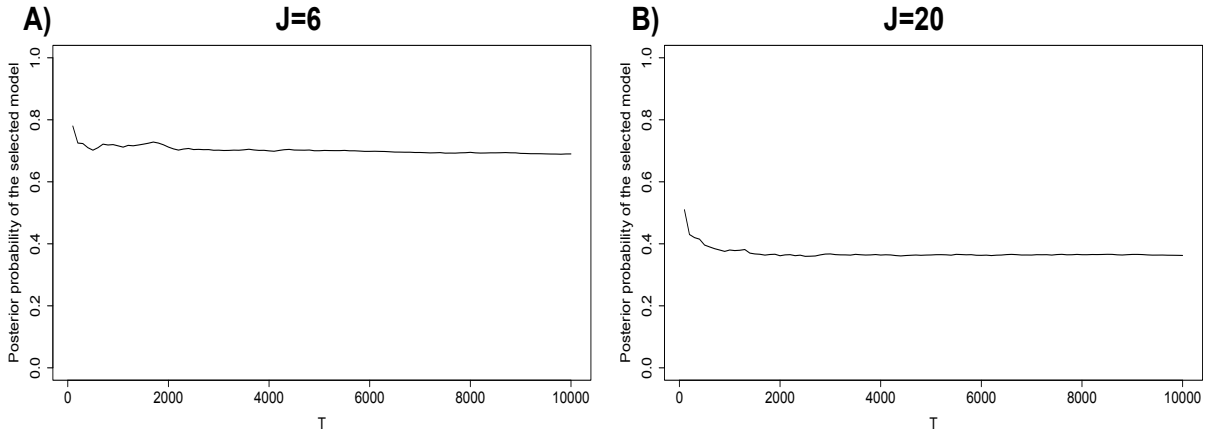

Figure 3: Convergence of Gibbs sampler chain when (A)$J = 6$ and (B)$J = 20$

Figure 3 shows the posterior probability of the selected model against the iteration number for γ given $J = 6$ and $J = 20$ under fixed prior probabilities. The chain starts with the null model $\gamma = 0$. Both, when $J = 6$ and $J = 20$, the chain converges rapidly, stabilizing after around 2,000 iterations. However, due to the time limitations of the study, we decided to discard the first 1,000 iterations for the variable selection, despite the Gibbs sampler chain suggesting the need for more iterations.

First, we explored how often the true model is selected when the sample size varied from $n = 50$ to $n = 500$ and the effect size varied from $\beta = .1$ to $\beta = 3$. Figure 4 plots the probability of selecting the true model (top panels) and posterior probability of the true model (bottom panels) from the 100 samples as a function of

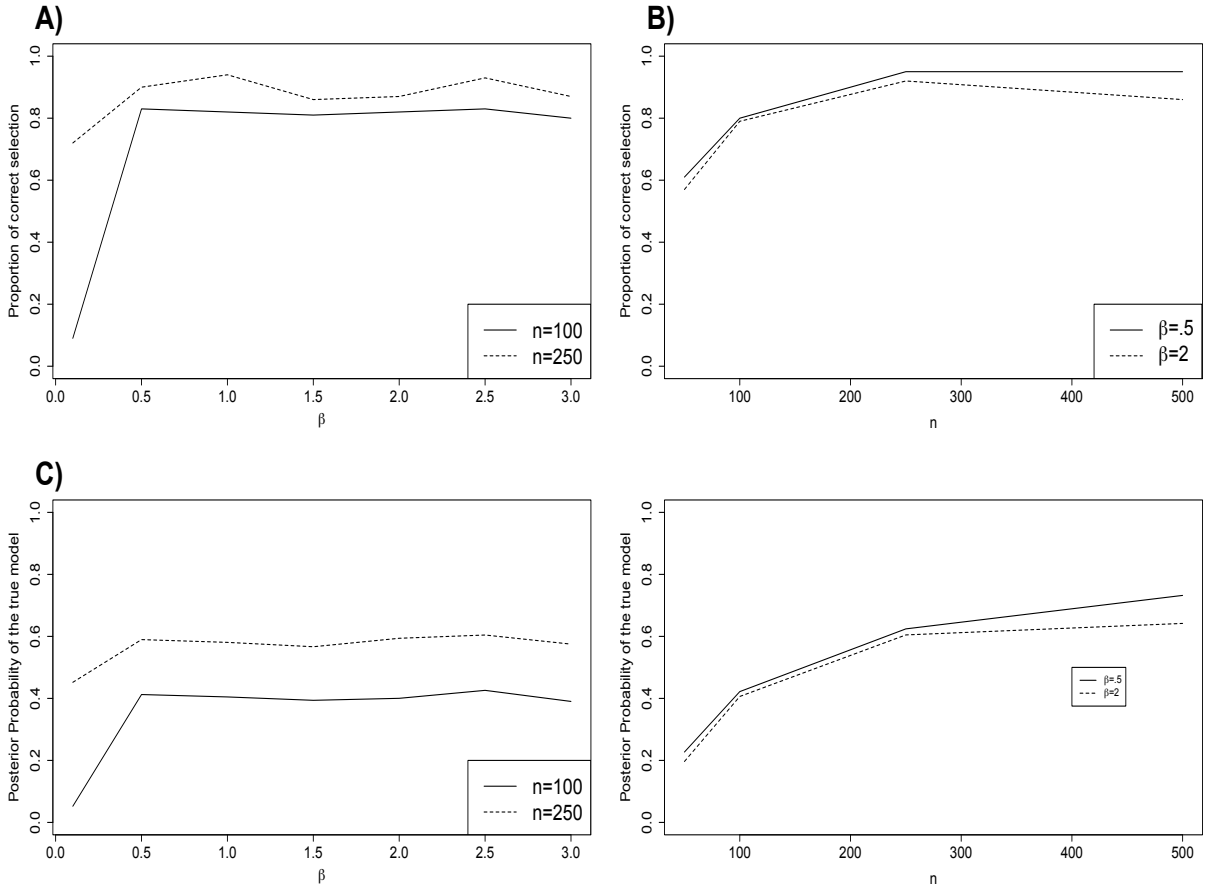

Figure 4: Proportion of selecting the true model (A & B) and posterior probability of the true model (C & D) given $J = 10$. In A and C , set $n = 100$ and $n = 250$, β varied from .1 to 3. In B and D, set $\beta = .5$ and $\beta = 2$, n varied from 50 to 500.

β given $n = 100$ and $n = 250$ as well as a function of n given $\beta = .5$ and $\beta = 2$. Both ratio of correct selection and the posterior probability of the true model increases as n grows up, which implies that the larger the sample size, the higher the chance to select the true model. The correct selection ratio and the posterior probability of the true model increase with larger effect until the point where $\beta = .5$ for both $n = 100$ and $n = 250$ but interestingly after that point both stabilize. Likewise, in the left panels, we can see that different β values do not increase the proportion of correct selection other than randomness.

Next, we explored how often the true model is selected when we set `k=4` and varied J from $J = 6$ to $J = 30$, and when we set $J = 10$ and varied the number of

knots from `k=4` to `k=12`.

	# of Variables (J)					# of Knots (`k`)				
	6	**10**	**15**	**20**	**30**	**4**	**6**	**8**	**10**	**12**
A	.478	.407	.365	.351	.300	.391	.479	.506	.445	.424
B	.83	.80	.79	.77	.74	.80	.90	.91	.86	.84
C	.59	**1**	1.53	2.08	3.37	**1**	1.074	1.135	1.226	1.438

Table 1: Posterior probabilities of the true model (A), and proportion of the correct selection (B) CPU time (C) given $n = 100$ and $\beta = .8$ across different values of `k` and J values.

Table 1 presents the posterior probabilities of the true model, the proportion of correct selections from the 100 samples and the central processing unit (CPU) time, as a function of J and `k` when $n = 100$ and $\beta = 0.8$. For CPU, the model with $J = 10$ and `k=4` was taken as the reference. As J increases, the posterior probability of the true model decreases because the model becomes more complex with fixed prior probabilities. However, the proportion of selecting the true model remains relatively stable. In other words, even as the posterior probability of the true model decreases in more complex models, the algorithm can still identify the true model.

As previously discussed, the precise selection of `k` is not typically crucial; it should be chosen to be large enough to capture the underlying nonlinear relationship, while also being small enough to ensure acceptable computational efficiency. Since, the nonlinear relationship considered in this study is not very complex/wiggly, we have set `k=4` to simplify the computation. The results show that increasing `k` to 6 (i.e., 5 basis functions) leads to an approximately 0.1 increase in the posterior probability of the true model, indicating a better capture of the nonlinear relationship compared to using 3 basis functions (Table 1). However, further increasing `k` results in only marginal improvements in the posterior probability. For instance, from `k=6` to `k=8`, there is only a slight increase in the posterior probability. While increasing `k` beyond a certain point might not lead to substantial improvements in the model's fit, it does increase computational burden. Therefore, it is essential to strike a balance between model complexity and computational efficiency when choosing the appropriate value of `k`. Furthermore, when `k` is greater 8, both the proportion of correct selection and

the posterior probability of the true model decrease. This suggests that adding more basis function does not enhance the fit but instead increases the penalty applied to nonlinear models. In other words, the algorithm penalizes the nonlinear models excessively when k is very high, causing the linear model to have a higher posterior probability even when the true relationship is nonlinear.

Next, we illustrate whether the selection algorithm is corrected for the multiplicity problem setting Dirichlet prior on, p_r (Section 4). Based on the sampling scenario in (17), all variables except for the first four are irrelevant to the outcome variable. Among these four relevant predictors, the first two exhibit nonlinear effects, while the remaining two demonstrate linear effects on the outcome variable. Consequently, we would expect that the number of included variables as both linear and nonlinear is consistently around 2. For each J, 100 samples were generated and the average number of identified effect types of the candidate variables across different J is plotted in Figure 5. From this figure, we can clearly observe that number of included variables as both linear and nonlinear effect types remains stable with increasing J. Therefore, it can be concluded that the Dirichlet prior on p_r performs very well in terms of controlling multiplicity when the number of candidate variables is large.

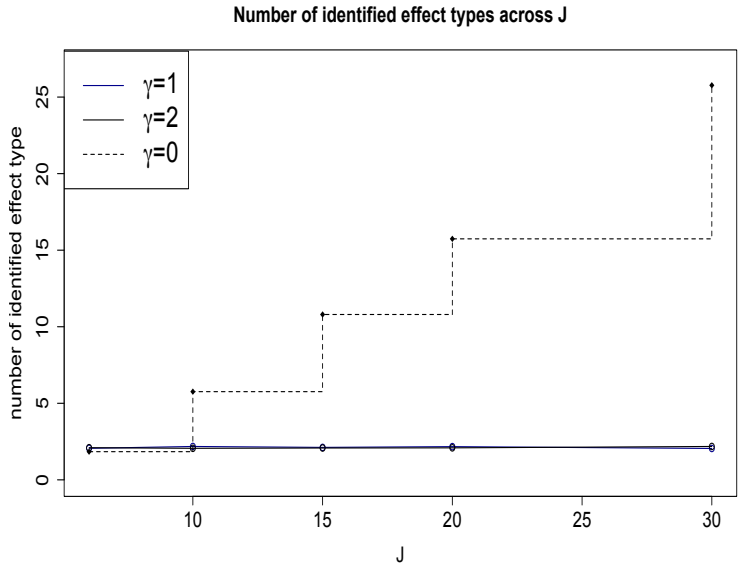

Figure 5: Number of each effect type included in the selected model across $J \in \{6, 10, 15, 20, 30\}$.

6. Empirical Examples

In this section two real data examples are used to illustrate the proposed variable selection. In particular we apply MCMC model search method to real datasets with relatively large number of candidate variables. To address potentially more complex nonlinear relationships, we set `k = 6` and performed 2,000 iterations for both examples. Details of the two datasets and the R code for the analysis can be found in **Appendix-C**.

I. Boston Housing Data

As the first example, we re-analyzed the Boston housing data described by Harrison and Rubinfeld (1978) and openly accessible in the R package "spdep" (Bivand, 2023). The data contains 506 observations and 20 variables in total. We removed the variables: TOWN, TOWNNO, TRACT, LON and LAT. The outcome variable is the median value of owner-occupied houses (in USD 1000's) with $J = 13$ candidate predictor variables including, e.g., crime rate, tax rate, pupil teacher ratios. Our target is to select a subset of variables along with their effect types to estimate the value of owner-occupied houses in Boston.

In the original analysis, certain variables were transformed using logarithmic or squared transformations to fit linear regression models. In this study, all the variables were analyzed in their original form without any transformation. The model $M_{\gamma'} = (1, 0, 0, 1, 2, 2, 0, 1, 1, 1, 1, 1, 2)$ got selected as the true model with a posterior probability of .378. Hence, our recommendation is to exclude the variables ZN, INDUS, and AGE from the model as they appear to have little or no influence on the outcome. Among the included variables, we observed that variables NOX, RM, and LSTAT exhibit nonlinear effects on the outcome variable.

II. Ozone Data

In this example, we explored Los Angeles ozone pollution data analyzed by Breiman and Friedman (1985) and available in the R package "mlbench" (Leisch & Dimitriadou, 2023). The dataset comprises daily measurements of maximum ozone concentration near Los Angeles and eight meteorological variables ($J = 8$). Our goal is to select a

subset of variables and their effect types to estimate the daily ozone concentration in Los Angeles.

Similar to the previous example, no transformations were applied to the data, and we considered the original form of the variables for analysis. The model $M_{\gamma'} = (1, 0, 1, 2, 1, 1, 0, 0)$ got selected as the true model with a posterior probability of .302. Thus, our recommendation is to exclude the variables "wind", "ibt" and "vis". Among the variables included in the model, we observed that temperature exhibits a nonlinear effect on the daily ozone concentration.

7. Conclusion

In this study, we presented a Bayesian variable selection scheme for both linear and nonlinear effects in regression models. When the number of candidate variables is large, MCMC model search method which only visits models with high posterior probabilities has been adopted for fast computation and multiplicity is controlled by setting Dirichlet prior on inclusion probabilities.

Compared to existing common variable selection methods, the straightforward advantage of the presented method that it gives information on the relevance of the candidate variables as well as whether the included variable has a linear or nonlinear effect on the outcome variable. Furthermore, it allows researchers to quantify the relative evidence in data for the selected model via the posterior probability. Although, prior specification and marginal likelihood computation limits the popularity of Bayesian models compared to those in the frequentist setting, we considered the selection problem in the context of GAMs and used BIC approximation of marginal likelihood. GAMs are useful to capture when the relationship is complex or not well understood so the algorithm can be used to identify a wide range of nonlinear relationships as well as the linear relationships. In this sense, the presented Bayesian variable selection method is very straightforward to use and apply to many different effect types as frequentist models, but with a quantification of the relative evidence in the data under the selected model.

We conducted a simulation study to evaluate the numerical behavior of the proposed Bayes factor and the MCMC model search in different setups. Several conclusions can be made from the results. First, the Bayes factor is able to identify the true model with the highest posterior probability among the model space across different settings. Second, the MCMC model search algorithm performs quite well in terms of fast convergence and accurate selection. Third, the specification of prior inclusion probabilities can effectively control for the multiplicity.

Due to the time constraint, however, we focused solely on exponential nonlinear relationships and omitted the burn-in at 1,000 iterations for the MCMC model search algorithm. While MCMC model search is recommended for cases with more than 15 candidate variables (Gu et. al, 2022; George & McCulloch, 1997), the examples used in our study involved fewer than 15 candidate variables to demonstrate the method's practicality. Lastly, we set a noninformative Dirichlet prior on the probabilities of observing each effect type by setting $p_r^0 = 1/3$ at $t = 0$. Different settings can be considered to assess the sensitivity of the model search method to the prior specification. Therefore, the model search algorithm can be further explored with more intricate simulation settings and larger datasets.

We assumed fixed prior model probabilities throughout the work, which does not allow to use prior information and it is waste of information if such an information exists on effect type of the candidate variables. In the MCMC algorithm, multiplicity is controlled for a large number of candidate variables via the Dirichlet prior, but it still assumes fixed prior model probabilities. In case of smaller number of candidate variables, multiplicity can still lead to incorrect conclusions. Thus, the proposed Bayes factor can be improved using prior model probabilities where researchers can incorporate prior knowledge rather than assuming all the models are equally likely to be observed. Finally, this paper does not discuss the high dimensional data case where 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.

References

[1] Bivand, R. (2023), "Package 'spdep'" R package version, 1.2-8.

[2] Edwards, W., Lindman, H. R., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. Psychological Review, 70(3), 193–242. https://doi.org/10.1037/h0044139

[3] F. Bach (2008). Exploring large feature spaces with hierarchical multiple kernel learning. In NIPS.

[4] F. Li, Y. Yang, and E. P. Xing (2005). From lasso regression to feature vector machine. In NIPS.

[5] 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

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

[7] Gu, X., Hoijtink, H., & Mulder, J. (2020). Bayesian One-Sided Variable selection. Multivariate Behavioral Research. https://doi.org/10.1080/00273171.2020.1813067

[8] Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. Statistical Science, 14(4), 382–401. http://www.jstor.org/stable/2676803 https://doi.org/10.1198/016214508000000337

[9] Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. Journal of the American Statistical Association, 90(430), 773–795. https://doi.org/10.2307/2291091

[10] Kass, R. E., & 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

[11] Kumar, A. K. (2016) Bayesian variable selection in linear and nonlinear models. [Unpublished doctoral dissertation]. Northern Illinois University.

[12] Leisch, F., & Dimitriadou, E. (2023). "Package 'mlbench'" R package version, 2.1-3.1.

[13] M. Yamada, W. Jitkrittum, L. Sigal, et al (2014). High dimensional feature selection by feature-wise kernelized lasso. Neural Comput, 26(1):185–207.

[14] Mulder, J. (2022). Bayesian testing of linear versus nonlinear effects using Gaussian process priors. The American Statistician, 77(1), 1–11. https://doi.org/10.1080/00031305.2022.2028675

[15] Navot, L. Shpigelman, N. Tishby, and E. Vaadia (2005). Nearest neighbor based feature selection for regression and its application to neural activity. In NIPS.

[16] Niu, Y. (2020, October 27). Bayesian Variable Selection in Multivariate Nonlinear Regression with Graph Structures. arXiv.org. https://arxiv.org/abs/2010.14638

[17] Park, T., & Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association, 103(482), 681–686.

[18] Peduzzi, P. N., Hardy, R. J., & Holford, T. R. (1980). A Stepwise Variable Selection Procedure for Nonlinear Regression Models. Biometrics, 36(3), 511–516. https://doi.org/10.2307/2530219

[19] Raftery, A. E. (1995). Bayesian Model Selection in Social Research. Sociological Methodology, 25, 111–163. https://doi.org/10.2307/271063

[20] Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). "Sparse Additive Models," Journal of the Royal Statistical Society, Ser. B, 71, 1009–1030.

[21] Rockova, 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

[22] 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

[23] Wagenmakers, E. (2007). A practical solution to the pervasive problems of p-values. Psychonomic Bulletin & Review, 14(5), 779–804. https://doi.org/10.3758/bf03194105

[24] Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society Series B-statistical Methodology, 65(1), 95–114. https://doi.org/10.1111/1467-9868.00374

[25] Wood, S.N. (2023), "Package 'mgcv'" R package version, 1.9 -0.

8. Appendix

The R-code including the simulation study and the application of the method to the real data examples;

- **Appendix-A:** Bayes Factor Consistency Check
- **Appendix-B:** MCMC Model Search Method Performance
- **Appendix-C:** Real Data Examples