Math. Appl. **11** (2022), 143–154
DOI: 10.13164/ma.2022.11

BAYESIAN VARIABLE SELECTION FOR LINEAR REGRESSION WITH THE κ-G PRIORS

ZICHEN MA AND ERNEST P. FOKOUÉ

Abstract. In this paper, we propose a method that balances between variable selection and variable shrinkage in linear regression. A diagonal matrix $\boldsymbol{G}$ is injected to the covariance matrix of prior distribution of the coefficient vector $\boldsymbol{\beta}$, with each g_j, bounded between 0 and 1, on the diagonal serving as a stabilizer of the corresponding β_j. Mathematically, a g_j value close to 0 indicates that the β_j is nonzero, and hence the corresponding variable should be selected, whereas the value of g_j close to 1 indicates otherwise. We prove this property under orthogonality. Computationally, the proposed method is easy to fit using automated programs such as JAGS. We provide three examples to verify the capability of this methodology in variable selection and shrinkage.

1. Introduction

Consider the Linear model given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right), \tag{1.1}$$

where $\boldsymbol{y}$ is an $n \times 1$ response vector, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p]$ an $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ a $p \times 1$ regression coefficient vector, and $\boldsymbol{\epsilon}$ the random error. Let $\mathcal{X}$ be the collection of all covariates in (1.1). In many applications where the size of $\mathcal{X}$ is large, only a small portion of covariates affect the response whereas the others are irrelevant [16]. This leads to the key question of identifying those "important" covariates. Approaches to answering this question can be loosely dichotomized into two general strategies, variable selection or variable shrinkage. In general, both seek the best fit that balances between maximizing the given data likelihood and minimizing the complexity of the model. The difference lies in how to achieve this goal. Variable selection concerns finding a subset of $\mathcal{X}$ which produces the best fit based on some criterion. An unselected covariate $\boldsymbol{x}_j$ has estimated coefficient $\widehat{\beta}_j$ in the final model. On the other hand, variable shrinkage, or regularization, techniques minimize the residual sum of squares associated with (1.1), subject to some constraints that penalize the complexity of the model. This leads to shrinking some estimated coefficients $\widehat{\beta}_j$ toward 0, but may not be exactly 0, faster than others.

Under the frequentist framework, stepwise regression has been widely used as a technique in variable selection [13]. There are several different ways to implement

MSC (2020): primary 65C20.
Keywords: Bayesian linear regression; variable selection; variable shrinkage; g-prior.

stepwise regression. For instance, in the procedure of forward selection, starting from the null model $\boldsymbol{y} = \boldsymbol{\mu}+\boldsymbol{\epsilon}$, where $\boldsymbol{\mu}$ is independent of $\boldsymbol{X}$, variables are added to the null model one at a time until a certain stopping rule is met. The stopping rule often utilizes an information criterion, such as AIC or BIC. The selected covariates in the final model are considered "important". [15] proposed a stepwise regression method that is suitable for high-dimensional sparse linear regression with $p \gg n$. An alternative to stepwise regression is the best subsets regression, which ranks linear models for all possible subsets of $\mathcal{X}$ according to a certain criterion, often the Mallows' C_p-statistic [19]. Models with values of C_p close to p are usually considered the best [11].

Ridge regression, due to [14], is one of the earliest practices of frequentist variable shrinkage in a linear model. The original idea is to boost the diagonal in the ill-posed matrix $\boldsymbol{X}'\boldsymbol{X}$ by adding a positive quantity λ so that the inversion is feasible. From the regularization point of view, ridge regression minimizes the residual sum of squares subject to the condition $||\boldsymbol{\beta}||_2^2 = \sum \beta_j^2$ less than some constant. The shrinkage parameter λ is simply the Lagrange multiplier in the constraint optimization. Increasing the value of λ results in the β_j's shrinking toward 0. However, ridge regression is not sparse as a result of the squared penalty. This problem is remedied in lasso [22], which alters the constraint to $||\boldsymbol{\beta}||_1 = \sum |\beta_j|$ less than some constant. Due to the L_1 penalty, the lasso solution produces $\widehat{\beta}_j = 0$ for some j, indicating explicitly that the corresponding covariate is irrelevant. Work along this line includes [23], [7], and [27].

Under the Bayesian framework, a widely used strategy in variable selection is as the following. Each covariate $\boldsymbol{x}_j \in \mathcal{X}$ is coupled with an indicator γ_j, which is equal to 1 if $\boldsymbol{x}_j$ is included in the model and 0 otherwise. Using this notation, every subset in $\mathcal{X}$ can be associated with an indicator vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$. Then, with certain priors on $\boldsymbol{\beta}$ and σ^2, the marginal distribution $p(\boldsymbol{y}|\boldsymbol{\gamma})$ under model $\boldsymbol{\gamma}$ can be computed by integrating out $\boldsymbol{\beta}$. Then, different models $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ can be compared using Bayes factor, given by $BF_{12} = [p(\boldsymbol{y}|\boldsymbol{\gamma}_1)/p(\boldsymbol{y}|\boldsymbol{\gamma}_2)]$. If model $\boldsymbol{\gamma}_1$ is better, the resulting Bayes factor is large. A thorough discussion on Bayes factor is given in [17]. An issue with this strategy is that the integration is often infeasible, leading to an intractable Bayes factor. [26] proposed an informative prior on $\boldsymbol{\beta}$ of the form $\boldsymbol{\beta}|\sigma^2 \sim N_p(\boldsymbol{0}, g\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ and an improper prior on σ^2 of the form $p(\sigma^2) \propto \sigma^{-2}$. This choice of prior distributions on $\boldsymbol{\beta}$ and σ^2 leads to a tractable Bayes factor, which makes it attractive in the variable selection. In practice, a reference model is usually chosen to be the full model $\boldsymbol{\gamma} = \boldsymbol{1} = (1, \ldots, 1)'$ or the null model $\boldsymbol{\gamma} = \boldsymbol{0} = (0, \ldots, 0)'$. Suppose the reference model is chosen to be the null model. For all other models, an associated Bayes factor can be computed by $BF = [p(\boldsymbol{y}|\boldsymbol{\gamma})/p(\boldsymbol{y}|\boldsymbol{0})]$. The best model has the largest Bayes factor.

Theoretically the method described above should exhaust all 2^p possible subsets of $\mathcal{X}$, including the null model and the full model. Difficulty quickly arises when the dimensionality increases, due to the fact that this method searches through the model space of size 2^p. Certain works have been done to resolve this issue. Most notably, [9] proposed an empirical method of stochastic search variable selection (SSVS). Each β_j is selected or rejected based on a Monte Carlo average of γ_j from a Gibbs sampler. The Monte Carlo average of γ_j is called the posterior inclusion

probability (PIP) of β_j. A large value in PIP_j implies that the corresponding estimated γ_j equals 1, β_j nonzero, and hence the variable is likely to be in the true model, which the authors named as "promising variable".

Similar work can be seen in [2], in which the authors proposed a median probability model rather than a highest probability model, and the variables are selected based on a criterion of $PIP_j > 0.5$. Further, [8] modified the method in [2] to a prevalence model, which solved the problem that such median probability model may not exist. Certain works have been done to summarize the Bayesian variable selection with the indicator method. [21] provides a thorough review of different methods in Bayesian variable selection. [12] gives a detailed comparison of different empirical Bayes methods, especially the Markov Chain Monte Carlo (MCMC) methods, regarding the Bayes factor.

We denote by $\boldsymbol{\beta_\gamma}$ the regression coefficient vector according to model $\boldsymbol{\gamma}$. Certain thoughts have been given to the prior of $\boldsymbol{\beta}$ and $\boldsymbol{\beta_\gamma}$ instead of the traditional g-prior. [10] provided a prior of $\boldsymbol{\beta_\gamma}$ of the form $\boldsymbol{\beta_\gamma} \sim N(\mathbf{0}, \boldsymbol{D_\gamma R_\gamma D_\gamma})$, where $\boldsymbol{D_\gamma}$ is a diagonal matrix and $\boldsymbol{R_\gamma}$ is symmetric. Such a prior gives a good generalization of g-prior. [1] gave an alternative that follows $\boldsymbol{\beta_\gamma} \sim N(\mathbf{0}, g\sigma^2(\boldsymbol{X}'_{\boldsymbol{\gamma}}\boldsymbol{A_\gamma}\boldsymbol{X_\gamma})^{-1})$, where $\boldsymbol{A_\gamma}$ is symmetric and weighs different observations, but not the covariates, and $\boldsymbol{X_\gamma}$ is the design matrix according to model $\boldsymbol{\gamma}$. Moreover, multiple works have been done to extend the original Zellner's g-prior. Notably, [18] proposed a study on mixtures of g-priors which provides a family of hyperpriors on g while still preserving the tractability on the marginal likelihood. [4] developed an extension of Zellner's g-prior to generalized linear models, given a large family of hyperpriors on g. [20] introduced a fully Bayes formulation with an orthogonal decomposition on the matrix $\boldsymbol{X}'_{\boldsymbol{\gamma}}\boldsymbol{X_\gamma}$, which resolves the issue of $p \ggg n$. All the works mentioned above rely on the indicator method, which is classic but somewhat redundant. At the worst, the methods still have to face the model space of size 2^p.

An alternative to the indicator method is through variable shrinkage. [24] introduced a method called the relevance vector machine (RVM). The prior on $\boldsymbol{\beta}$ is given by $\beta_j \overset{ind}{\sim} N(0, \alpha_j^{-1})$ for $j = 1, \ldots, p$. The parameter α_j serves as a stabilizer. That is, since the coefficient β_j is a priori centered at 0, the prior variance approaches 0 as $\alpha_j \to \infty$. On the other hand, the prior becomes flat as $\alpha_j \to 0$. Interestingly, as stated in [25], combining the non-sparse normal prior on $\boldsymbol{\beta}$ with a gamma hyperprior on each of the α_j's, the marginal of $\boldsymbol{\beta}$ becomes a multivariate t-distribution after integrating out the α_j's, which leads the RVM to a sparse selection machine. As a side note, this property of sparsity is even more elegant when the input in the linear model is raised from feature space to kernel space, which is the main focus in [24, 25], but not in our work.

Further, a global-local shrinkage technique that has gained much attention in recent years is the horseshoe priors, due to [6]. Comparing to RVM, in addition to assigning each β_j its own prior variance, it formulates the prior of β_j as $\beta_j \sim N(0, \lambda_j^2\tau^2)$ with $\lambda_j \sim C^+(0, 1)$, a standard half Cauchy distribution. The global parameter τ shrinks all β_j's to 0, while the local parameter λ_j allows the specific β_j to escape from the shrinkage. For a detailed exposition on the horseshoe priors, its relation to lasso, and numerical examples, see [3].

Standard variable selection techniques usually have hard thresholds, meaning that a covariate is either included or excluded from the model, whereas standard variable shrinkage methods at times do not provide explicit information about whether a variable is irrelevant. In this article, we propose a method from the Bayesian perspective that, on the one hand, provides a soft threshold for variable selection, while, on the other hand, shrinks the irrelevant coefficients to a certain degree. Section 2 provides a thorough discussion on the formulation of the proposed method. Section 3 provides three examples that demonstrate the capability of the proposed methodology in variable selection. Of the three examples, the first two are simulated examples taken from [9], while the third one is an application on a real data set. And finally, we provide a conclusion in Section 4.

2. Method

Section 2.1 details the proposed hierarchical model and derives the posterior distributions of the regression coefficients. Section 2.2 discusses the posterior distribution of the shrinkage parameter and its usage in variable shrinkage. Specifically, an important result is given, which links the behavior of the shrinkage parameters and the "significance" of corresponding covariates under orthogonal design matrix. Section 2.3 provides the posterior distribution of the scale parameters. To avoid confusion, we adopt the phrase promising variable from [9].

2.1. Formulation of the κ-$\boldsymbol{G}$ model

Consider the linear model in (1.1). The idea of the proposed method is to construct an informative prior on $\boldsymbol{\beta}$, in which, for $j = 1, \ldots, p$, the covariance term couples each covariate $\boldsymbol{x}_j$ with a hyperparameter g_j with continuous support over $(0, 1)$. Given hyperpriors on each g_j, the relationship between the covaraite $\boldsymbol{x}_j$ and the response $\boldsymbol{y}$ is reflected through the posterior of g_j. Formally, let $\boldsymbol{G} = diag(g_1, \ldots, g_p)$ be a p-dimensional diagonal matrix. The proposed hierarchical model is given by

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 &\sim N_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right) \\ \boldsymbol{\beta}|\kappa, \sigma^2 &\sim N_p\left(\boldsymbol{0}, \kappa\sigma^2\left[\boldsymbol{G}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{G}\right]^{-1}\right) \\ g_1, g_2, \ldots, g_p &\overset{iid}{\sim} \text{Beta}(a, a) \\ \kappa^{-1} &\sim \text{Gamma}(\alpha, \theta) \\ p(\sigma^2) &\propto \sigma^{-2}. \end{aligned} \tag{2.1}$$

Consider the intuition of this model when the design matrix is orthogonal; i.e. $\boldsymbol{x}_{j_1}'\boldsymbol{x}_{j_2} = 0$ for all $j_1 \neq j_2$. Provided κ and σ^2 are non-zero, the prior on β_j has infinite variance if g_j is strictly 0. On the other hand, the prior variance of β_j is the same as that under Zellner's g-prior, which is the inverse Fisher information scaled by κ, when g_j is strictly 1. We let g_j vary between 0 and 1 with a beta prior distribution symmetric about $\frac{1}{2}$. The parameter κ has the same practical meaning as the parameter g in Zellner's g-prior, serving as a global shrinkage parameter. A larger value of κ corresponds to more prior variability on $\boldsymbol{\beta}$. The gamma prior distribution on κ^{-1} is out of consideration for conjugacy. The use of Jeffrey's prior on σ^2 follows from [26].

Following the formulation in (2.1), the posterior of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}|\boldsymbol{G},\boldsymbol{y},\kappa,\sigma^2 \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \tag{2.2}$$

where the posterior mean and variance are given by

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}+\kappa^{-1}\boldsymbol{G}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{G}\right)^{-1}\boldsymbol{X}'\boldsymbol{y},\\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}+\kappa^{-1}\boldsymbol{G}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{G}\right)^{-1}.\end{aligned} \tag{2.3}$$

Denote by $\widehat{\boldsymbol{\beta}}^{(OLS)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$ the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$. The proposition below follows from (2.2) and (2.3).

Theorem 2.1. *Let $\kappa > 0$. We have*

(i) *The posterior of $\boldsymbol{\beta}$ approaches the OLS estimator as $g_j \to 0$ for all $j = 1,\ldots,p$ with $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(OLS)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = Var(\widehat{\boldsymbol{\beta}}^{(OLS)})$.*

(ii) *The posterior of $\boldsymbol{\beta}$ approaches the posterior using Zellner's g-prior as $g_j \to 1$ for all $j = 1,\ldots,p$, with $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \frac{\kappa}{\kappa+1}\cdot\widehat{\boldsymbol{\beta}}^{(OLS)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \frac{\kappa}{\kappa+1}\cdot Var(\widehat{\boldsymbol{\beta}}^{(OLS)})$.*

The results above are immediate by setting $\boldsymbol{G}$ to be the zero matrix $\boldsymbol{0}_{p\times p}$ in (i) and the identity matrix $\boldsymbol{I}_{p\times p}$ in (ii), respectively.

In the first part of Proposition 2.1, observe that $g_j \to 0$ for all $j = 1,\ldots,p$ is equivalent to assigning a flat prior on all β_j's, which provides mininal amount of a priori information on $\boldsymbol{\beta}$. As a result, the posterior mean coincides with the OLS estimator, which is also the frequentist estimator of $\boldsymbol{\beta}$ under maximum likelihood. On the other hand, in the second part of the proposition, if $g_j \to 1$ for all $j = 1,\ldots,p$, the prior precision is proportional to the Fisher information matrix of $\boldsymbol{\beta}$, and the scalar κ takes on the role of Zellner's g. The prior on $\boldsymbol{\beta}$ with $\boldsymbol{G} = \boldsymbol{I}$ can be viewed as providing the most amount of prior information. Overall, the parameter g_j attempts to balance in between the two extremes and to provide a reasonable amount of information on $\boldsymbol{\beta}$.

2.2. Posterior distribution of $\boldsymbol{G}$

We now examine the posterior distribution of $\boldsymbol{G}$ and its usage as a shrinkage parameter. Integrating out $\boldsymbol{\beta}$, the posterior $\boldsymbol{G}$ is given by

$$\begin{aligned}\pi(\boldsymbol{G}|\boldsymbol{y},\kappa,\sigma^2) &\propto \pi(\boldsymbol{G})\int\left[f(\boldsymbol{y}|\boldsymbol{\beta},\kappa,\sigma^2)\times\pi(\boldsymbol{\beta}|\boldsymbol{G},\kappa,\sigma^2)\right]\mathrm{d}\boldsymbol{\beta}\\ &\propto |\boldsymbol{G}|^{a}|\boldsymbol{I}-\boldsymbol{G}|^{a-1}|\boldsymbol{X}'\boldsymbol{X}+\kappa^{-1}\boldsymbol{G}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{G}|^{-1/2}\times\\ &\quad\exp\left[\frac{\boldsymbol{y}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}+\kappa^{-1}\boldsymbol{G}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{G}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}}{2\sigma^2}\right],\end{aligned} \tag{2.4}$$

where a represents the hyperparameter in the *iid* Beta(a,a) hyperprior on each g_j in (2.1). The intractability creates difficulty in the discussion on the function of $\boldsymbol{G}$, much of which involves inverting the matrix $\left(\boldsymbol{X}'\boldsymbol{X}+\kappa^{-1}\boldsymbol{G}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{G}\right)$. In this article, we do not pursue the general case in (2.4). Instead, we proceed by showing (i) an example with $p = 2$ as an intuitive illustration; and more formally (ii) a discussion on the usage of $\boldsymbol{G}$ as a shrinkage parameter when the design matrix is orthogonal.

For the purpose of illustration, consider a linear model of $\boldsymbol{y} = 1 \cdot \boldsymbol{x}_1 + 0 \cdot \boldsymbol{x}_2 + \boldsymbol{\epsilon}$, where the random error is given by $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I})$. In other words, the response $\boldsymbol{y}$ depends on the promising variable $\boldsymbol{x}_1$ but not the unpromising variable $\boldsymbol{x}_2$. Figure 1 provides an illustration of $\pi(g_1, g_2|\boldsymbol{y})$ assuming $\kappa = \sigma^2 = 1$. The promising covariate $\boldsymbol{x}_1$ is associated with g_1, and the unpromising covariate $\boldsymbol{x}_2$ is associated with g_2. Note from the figure that the posterior is maximized at the boundary of g_2 with $g_2 \to 1$, and g_1 close to 0. This observation sheds some light on how the posterior of $\boldsymbol{G}$ is related to variable shrinkage and selection.

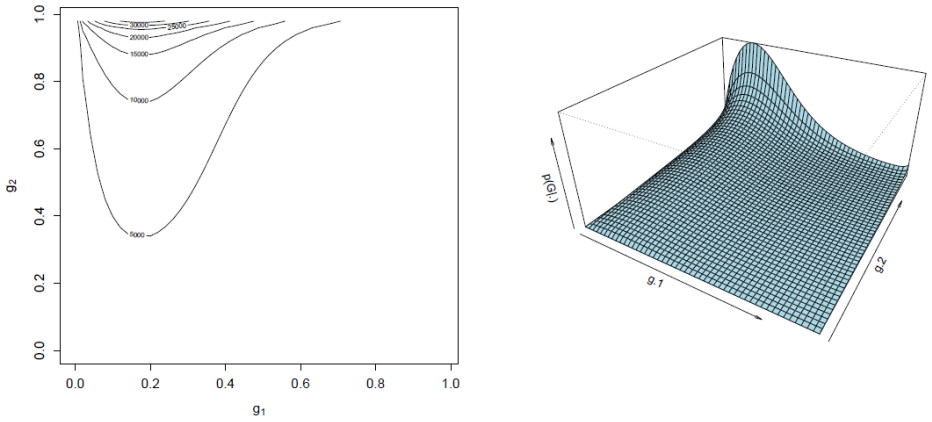

Figure 1. Contour (left) and perspective (right) plot of the posterior of $\boldsymbol{G}$ with $p = 2$. The response is given by $\boldsymbol{y} = 1 \cdot \boldsymbol{x}_1 + 0 \cdot \boldsymbol{x}_2 + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Assume $\kappa = \sigma^2 = 1$.

Consider the situation under orthogonal design matrix, where

$$\boldsymbol{X}'\boldsymbol{X} = diag(\boldsymbol{x}_1'\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p'\boldsymbol{x}_p).$$

In this case, (2.4) is simplified to

$$\begin{aligned} \pi(\boldsymbol{G}|\boldsymbol{y}, \kappa, \sigma^2) &= \prod_{j=1}^{p} \pi(g_j|\boldsymbol{y}, \kappa, \sigma^2) \\ &\propto \prod_{j=1}^{p} g_j^a (1-g_j)^{a-1} (\kappa + g_j^2)^{-1/2} \exp\left[\frac{\kappa(\boldsymbol{x}_j'\boldsymbol{y})^2}{2\sigma^2 \boldsymbol{x}_j'\boldsymbol{x}_j(\kappa + g_j^2)}\right]. \end{aligned} \tag{2.5}$$

The factorization indicates that the g_j's are *a posteriori* independent. Studying each $\pi(g_j|\boldsymbol{y}, \kappa, \sigma^2)$ from the perspective of the maximum a posteriori (MAP) estimator, the property of the posterior of g_j is summarized in the following proposition.

Theorem 2.2. *A promising covariate $\boldsymbol{x}_j$ has a corresponding g_j close to 0, where as an unpromising covariate has a corresponding g_j close to 1.*

Proof. We give a somewhat heuristic proof here. Without loss of generality, assume $a = \kappa = \sigma^2 = 1$. Then, the posterior density of each g_j in (2.5) reduces to

$$\begin{aligned}\pi(g_j|\boldsymbol{y}) &\propto g_j(1+g_j^2)^{-1/2}\exp\left[\frac{(\boldsymbol{x}_j'\boldsymbol{y})^2}{2\boldsymbol{x}_j'\boldsymbol{x}_j(1+g_j^2)}\right] \\ &= g_j(1+g_j^2)^{-1/2}\exp\left[\frac{||\boldsymbol{y}||^2\cos^2\theta_j}{2(1+g_j^2)}\right],\end{aligned} \tag{2.6}$$

where $||\boldsymbol{y}||^2 = \boldsymbol{y}'\boldsymbol{y}$ and θ_j is the angle between $\boldsymbol{x}_j$ and $\boldsymbol{y}$.

Unpromising covariate. For an unpromising covariate $\boldsymbol{x}_j$, it is reasonable to assume that $\cos\theta_j = 0$. Thus, (2.6) further simplifies to

$$\pi(g_j|\boldsymbol{y}) \propto g_j(1+g_j^2)^{-1/2},$$

which is continous and monotone increasing over $(0,1)$. Therefore, the MAP estimator of g_j for an unpromising covariate is given by $\widehat{g}_j = \arg\max_{g_j}\pi(g_j|\boldsymbol{y}) = 1^-$.

Promising covariate. For a promising covariate $\boldsymbol{x}_j$, $\cos^2\theta_j > 0$. Since all the terms on the exponent in (2.6) are positive, $\exp(\cdot)$ is a decreasing function of g_j on $(0,1)$. Therefore, in this case, the posterior of g_j is a compromise between the monotone increasing function $g_j(1+g_j^2)^{-1/2}$ and the monotone decreasing exponential function. With a moderately large $\|\boldsymbol{y}\|$, the decreasing exponential function becomes the dominant term in (2.6). Since $\|\boldsymbol{y}\|^2 = \sum_{i=1}^n y_i^2 \to \infty$ as $n \to \infty$ except for the trivial case of $\boldsymbol{y} = \boldsymbol{0}$, the MAP $\widehat{g}_j \to 0$ as $n \to \infty$ for the promising covariate. □

2.3. Posterior distribution of κ and σ^2

Lastly, we present the posterior distribution of κ and σ^2. By conjugacy, the posterior of κ is given by

$$\kappa^{-1}|\boldsymbol{y},\sigma^2,\boldsymbol{G} \sim \text{Gamma}(\widetilde{\alpha},\widetilde{\theta}),$$

where

$$\begin{aligned}\widetilde{\alpha} &= \frac{p}{2}+\alpha \\ \widetilde{\theta} &= \frac{\boldsymbol{\beta}'\boldsymbol{G}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{G}\boldsymbol{\beta}}{2\sigma^2}+\theta.\end{aligned}$$

Likewise, the posterior of σ^2 also has a closed-form expression, given by

$$\sigma^{-2}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{G},\kappa \sim \text{Gamma}\left(\frac{n+p}{2},\frac{s^2}{2}+\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{2}+\frac{\boldsymbol{\beta}'\boldsymbol{G}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{G}\boldsymbol{\beta}}{2\kappa}\right),$$

where

$$s^2 = \left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(OLS)}\right)'\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(OLS)}\right).$$

3. Numerical examples

In this section, we demonstrate the capability of our proposed method in variable selection through three examples. The first two examples are simulations taken from [9], while the third example is real application to the well-known `prestige` data set. For each example, we provide both the results based on the proposed method and a comparison to PIP using the traditional indicator-based method. The PIPs are computed using the BMS package in R.

The proposed model is fitted using the automated MCMC package JAGS in R. Our primary concern in writing the JAGS code is to elicit prior distributions for each g_j and for κ. Since we do not possess a priori knowledge of which covariates would be promising, the prior distribution on g_j is simply $g_j \sim \text{Beta}(1,1)$, which is equivalent to a standard uniform distribution over $(0,1)$. Similarly, the prior distribution of κ is given by $\kappa^{-1} \sim \Gamma(\alpha = 10^{-3}, \theta = 10^{-3})$, where α and θ are the shape and the rate parameters in the gamma distribution, respectively. This yields a proper, but fairly flat prior over the support of κ.

Example 3.1. Consider a linear regression with $p = 5$ predictors, each of length $n = 60$, given by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5 \sim N_{60}(\boldsymbol{0}, \boldsymbol{I})$. The response is

$$\boldsymbol{y} = \boldsymbol{x}_4 + 1.2\boldsymbol{x}_5 + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N_{60}(\boldsymbol{0}, 2.5^2\boldsymbol{I}),$$

where the "true" regression coefficient vector is $\boldsymbol{\beta} = (0,0,0,1,1.2)'$.

Table 1. Posterior median of g_j's under the κ-$\boldsymbol{G}$ method and the posterior inclusion probability (PIP) of [9] in Example 3.1.

$\boldsymbol{x}_1$-$\boldsymbol{x}_3$		$\boldsymbol{x}_4$-$\boldsymbol{x}_5$	
κ-$\boldsymbol{G}$	PIP	κ-$\boldsymbol{G}$	PIP
0.788	0.204	**0.071**	1.000
0.775	0.261	**0.056**	0.966
0.779	0.537		

The results are presented in Table 1. Since the posterior distribution of g_j is skewed, we use the posterior median $\widehat{g}_j$ as the point estimate. Note that small $\widehat{g}_j$'s correspond to PIPs close to 1, indicating that the associated covariates are promising. On the other hand, large $\widehat{g}_j$'s correspond to low PIPs, indicating that such covariates are unpromising. This result verifies Proposition 2.2.

Example 3.2. The second example involves $p = 60$ predictors, each of length $n = 120$, which exhibit moderate correlation. The purpose of this example, as was well-stated in [9], is to "demonstrate the practical potential (of the proposed method) for data sets involving many potential predictors".

For $j = 1, 2, \ldots, 60$, let $\boldsymbol{x}_j = \boldsymbol{x}_j^* + \boldsymbol{z}$, where $\boldsymbol{x}_j^* \overset{iid}{\sim} N_{120}(\boldsymbol{0}, \boldsymbol{I})$ independent of $\boldsymbol{z} \sim N_{120}(\boldsymbol{0}, \boldsymbol{I})$. This induces correlation of about 0.5 among all $\boldsymbol{x}_j$'s. The response is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N_{120}(\boldsymbol{0}, 2^2\boldsymbol{I}),$$

where $\boldsymbol{\beta}$ is set to be $(\beta_1, \ldots, \beta_{15})' = \mathbf{0}$, $(\beta_{16}, \ldots, \beta_{30})' = \mathbf{1}$, $(\beta_{31}, \ldots, \beta_{45})' = 2 \cdot \mathbf{1}$, and $(\beta_{46}, \ldots, \beta_{60})' = 3 \cdot \mathbf{1}$.

Table 2. Posterior median of g_j's under the κ-$\boldsymbol{G}$ method and the posterior inclusion probability (PIP) of [9] in Example 3.2.

$\boldsymbol{x}_1$-$\boldsymbol{x}_{15}$		$\boldsymbol{x}_{16}$-$\boldsymbol{x}_{30}$		$\boldsymbol{x}_{31}$-$\boldsymbol{x}_{45}$		$\boldsymbol{x}_{46}$-$\boldsymbol{x}_{60}$	
κ-$\boldsymbol{G}$	PIP	κ-$\boldsymbol{G}$	PIP	κ-$\boldsymbol{G}$	PIP	κ-$\boldsymbol{G}$	PIP
0.901	0.066	**0.802**	0.210	0.064	0.887	0.039	0.872
0.906	0.278	0.089	0.563	0.051	0.478	0.031	0.987
0.808	0.273	**0.779**	0.028	0.061	0.065	0.034	1.000
0.894	0.052	0.093	0.528	0.069	0.626	0.052	0.400
0.868	0.044	**0.923**	0.172	0.045	0.950	0.027	1.000
0.767	0.167	0.132	0.135	0.070	0.109	0.032	0.954
0.860	0.203	0.218	0.173	0.043	0.807	0.042	0.941
0.885	0.036	0.122	0.292	0.065	0.260	0.036	1.000
0.875	0.129	0.211	0.012	0.047	0.293	0.039	0.898
0.846	0.094	0.431	0.452	0.091	0.143	0.036	0.110
0.919	0.043	0.118	0.096	0.058	0.173	0.047	0.798
0.825	0.149	**0.794**	0.096	0.158	0.047	0.039	0.791
0.879	0.060	0.080	0.391	0.056	0.597	0.043	0.957
0.876	0.229	0.138	0.221	0.049	0.564	0.035	0.981
0.916	0.166	0.180	0.108	0.061	0.093	0.051	0.951

The comparison results between the posterior median $\widehat{g}_j$'s of the proposed method and PIPs in [9] are provided in Table 2. The four columns correspond to the four different values of regression coefficients. If one applies a simplistic decision rule such that a variable is included if $\widetilde{g}_j < 0.5$, the only false decisions under the proposed method of this paper are $\widetilde{g}_{16}$, $\widetilde{g}_{18}$, $\widetilde{g}_{20}$, and $\widetilde{g}_{27}$, which are identified as bold in Table 2. On the other hand, if we apply the simplistic rule that a covariate is deemed promising if the corresponding PIP is greater than 0.5, then almost all covariates between $\boldsymbol{x}_{15}$ and $\boldsymbol{x}_{30}$ and more than half of the covariates between $\boldsymbol{x}_{31}$ and $\boldsymbol{x}_{45}$ are unpromising, which contradicts the set-up of the example. Clearly, within the context of this moderately complex example, the proposed method is superior than the PIP method in [9].

Further, it is also of interest to compare the estimates of $\boldsymbol{\beta}$ obtained from the $\kappa - \boldsymbol{G}$ method and from the OLS. The results are presented in Table 3. The estimates of $\boldsymbol{\beta}$ under the proposed method are the posterior means of each β_j. First note that the Bayesian estimates and the OLS estimates for β_1 to β_{15} do not always agree on the direction of the coefficient. However, this is not of great concern since these variables are unpromising based on the results from Table 2. Moreover, as the true magnitude of β_j increases, the differences between the Bayesian estimates and the OLS estimates vaguely decrease.

Table 3. Posterior mean under the κ-$\boldsymbol{G}$ method and the OLS estimator of regression coefficients β_j's in Example 3.2.

β_1-β_{15}		β_{16}-β_{30}		β_{31}-β_{45}		β_{46}-β_{60}	
κ-$\boldsymbol{G}$	OLS	κ-$\boldsymbol{G}$	OLS	κ-$\boldsymbol{G}$	OLS	κ-$\boldsymbol{G}$	OLS
-0.522	-0.243	**-0.041**	1.154	2.004	1.806	2.991	3.098
-0.475	0.110	1.695	1.334	2.937	2.059	3.294	3.225
-0.036	0.266	**-0.083**	0.870	2.274	1.692	3.293	3.121
-0.245	-0.233	2.120	1.383	2.761	1.842	2.557	2.469
-0.351	0.092	**-0.411**	0.667	3.173	2.355	3.786	3.285
0.060	0.062	0.929	0.815	2.272	1.973	3.903	3.169
-0.255	0.359	0.682	1.173	3.863	2.615	3.041	2.446
-0.379	0.232	1.316	1.201	2.082	1.737	3.394	3.250
-0.575	-0.385	0.553	0.925	2.022	2.150	3.534	3.241
-0.755	-0.552	0.306	0.932	1.146	2.096	2.899	2.893
-0.694	-0.120	1.107	1.005	1.966	2.006	3.185	2.617
-0.112	-0.036	**0.026**	0.607	0.631	1.851	3.648	3.170
-0.260	-0.199	1.342	1.152	2.334	2.267	3.420	2.978
-0.431	0.029	1.244	1.262	2.333	1.895	3.640	3.113
-0.437	-0.242	0.729	1.151	1.527	1.569	4.011	3.232

Example 3.3. The third example is an application of the proposed method to the `Prestige` data set in the R library `car`. The data were collected from the mid-1960s to early 1970s [5]. The data consists of $n = 102$ different occupations, and the prestige of each occupation is regarded as response and regressed onto three predictors: average education of occupational incumbents ($\boldsymbol{x}_1$), average income of incumbents ($\boldsymbol{x}_2$), and percentage of incumbents who are women ($\boldsymbol{x}_3$).

Table 4. Posterior median of g_j's and posterior mean of β_j's in Example 3.3, with comparison to PIP and OLS.

	$\widetilde{g}$	PIP	$\widehat{\beta}$	$\widehat{\beta}^{(OLS)}$
$\boldsymbol{x}_1$	0.017	1.000	3.644	3.553
$\boldsymbol{x}_2$	0.063	1.000	1.284	1.388
$\boldsymbol{x}_3$	0.808	0.236	-0.022	-0.013

Posterior estimates for each g_j and β_j corresponding to the three predictors are given in Table 4. As was seen before, the posterior median of each g_j agrees with the posterior inclusion probability, while the posterior mean of β_j is reasonably close to the corresponding OLS estimate.

4. Conclusion

In this paper, we have demonstrated a new method for Bayesian variable selection in linear model that is completely independent of the traditional indicator variable method. The coefficient vector $\boldsymbol{\beta}$ is given a normal prior of the form $N(\mathbf{0}, \kappa\sigma^2(\boldsymbol{GX'XG})^{-1})$. By injecting the diagonal matrix $\boldsymbol{G}$ to the variance of the prior, each diagonal element g_j in $\boldsymbol{G}$ serves as a variance stabilizer such that the promising variables are selected based on the g_j's that are close to 0. Mathematically, when the covariates are orthogonal to each other, the g_j's are a posteriori independent. Within the support of $(0, 1)$, g_j is maximized toward 0 if the associated covariate is promising, and toward 1 if the associated covariate is unpromising. A posterior point estimator, e.g. the posterior median, can then be used as a soft threshold indicating the importance of the corresponding predictor $\boldsymbol{x}_j$. Computationally, this proposed hierarchical model can be readily fitted using JAGS.

In Section 3, we have demonstrated the usefulness of this methodology through three examples. The first and third example showed that the proposed methodology is capable of yielding correct results when the dimensionality is fairly low. Moreover, in the second example, we have demonstrated the competence of this new method not only under mildly large dimensionality, but also under moderate correlation among the predictors. In fact, we have shown that under such circumstances, the proposed method is able to provide even more compelling results than the traditional indicator variable method.

The proposed methodology possesses great potential for future works. From the theoretical aspect, theoretical results need to be developed when the predictors are not necessarily orthogonal. The difficulty in this task involves inverting the matrix $\left(\boldsymbol{X'X} + \kappa^{-1}\boldsymbol{GX'XG}\right)$. Further, in this paper, we have implicitly assumed the response and the predictors are all continuous. This restriction can certainly be extended to binary predictors, binary responses, or generalized linear regression in general.

References

[1] A. Aglieri and C. Parisetti, *A-g reference informative prior: a note on Zellner's g-prior*, J. R. Stat. Soc., Ser. D **37** (1998), 271–275.

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

[3] A.Bhadra, J. Datta, N. Polson and B. Willard, *Lasso meets horseshoe: A survey*, Statist. Sci. **34** (2019), 405–427.

[4] D. Bové and L. Held, *Hyper g-priors for generalized linear models*, Bayesian Anal. **6** (2011), 387–410.

[5] Statistics Canada, *1971 Census of Canada*, `https://publications.gc.ca/site/eng/9.834259/publication.html`.

[6] C. Carvalho, N. Polson and J. Scott, *Handling sparsity via the horseshoe*, J. Mach. Learn. Res. **5** (2009), 73–80.

[7] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, *Least angle regression*, Ann. Statist. **32** (2004), 407–451.

[8] E. Fokoué, *Estimation of atom prevalence for optimal prediction*, Contemp. Math. **447** (2007), 103–129.

[9] E. I. George and R. E. McCulloch, *Variable selection via Gibbs sampling*, J. Amer. Statist. Assoc. **88** (1993), 881–889.
[10] E. I. George and R E. McCulloch, *Approaches for Bayesian variable selection*, Stat. Sin. **7** (1997), 339–373.
[11] S. Gilmour, *The interpretation of Mallows' C_p-statistic*, J. R. Stat. Soc., Ser. D **45** (1996), 49–56.
[12] C. Han and B. Carlin, *Markov chain Monte Carlo methods for computing Bayes factor: A comparative review*, J. Amer. Statist. Assoc. **96** (2001), 1122–1132.
[13] R. Hocking, *The analysis and selection of variables in linear regression*, Biometrics **32** (1976), 1–49.
[14] A. Hoerl and R. Kennard, *Ridge regression: biased estimation for nonorthogonal problems*, Technometrics **12** (1970), 55–67.
[15] C.-K. Ing and T.-L. Lai, *A stepwise regression method and consistent model selection for high-dimensional sparse linear models*, Stat. Sin. **21** (2011), 1473–1513.
[16] W. Jeffreys and J. Berger, *Sharpening Ockham's razor on a Bayesian strop*, Technical report, University of Texas at Austin and Purdue University, 1991.
[17] R. Kass and A. Raftery, *Bayes factor*, J. Amer. Statist. Assoc. **90** (1995), 773–795.
[18] F. Liang, R. Paulo, G. Molina, M. Clyde and J. Berger, *Mixtures of g-priors for Bayesian variable selection*, J. Amer. Statist. Assoc. **103** (2008), 410–423.
[19] C. Mallows, *Some comments on C_p*, Technometrics **15** (1973), 661–675.
[20] Y. Maruyama and E. George, *Fully Bayes factors with a generalized g-prior*, Ann. Statist. **39** (2011), 2740–2765.
[21] R. O'Hara and M. Sillanpää, *A review of Bayesian variable selection: what, how, and which*, Bayesian Anal. **4** (2011), 85–118.
[22] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc., Ser. B **58** (1996), 267–288.
[23] R. Tibshirani, *The lasso method for variable selection in the Cox model*, Stat. Med. **16** (1997), 385–395.
[24] M. Tipping, *Sparse Bayesian learning and the relevance vector machine*, J. Mach. Learn. Res. **1** (2001), 211–244.
[25] M. Tipping, *Bayesian inference: an introduction to principles and practice in machine learning*, in: O. Bousquet, U. von Luxburg and G. Rätsch (eds.), *Advanced Lectures on Machine Learning*, Springer, 2004, pp. 41–62.
[26] A. Zellner, *On assessing prior distributions and Bayesian regression analysis with g-prior distributions*, in: P. K. Goel and A. Zellner (eds.), Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, Studies in Bayesian Econometrics and Statistics **6**, North-Holland, Amsterdam, 1986, pp. 233-243.
[27] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc., Ser. B **67** (2005), 301–320.

Zichen Ma, Department of Mathematics, Colgate University, Hamilton, NY USA
e-mail: zma@colgate.edu

Ernest P. Fokoué, School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY USA
e-mail: epfeqa@rit.edu