Understanding and assessment of a two-component G-prior in variable selection

Farnaz Solatikia

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

This Thesis is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
UNDERSTANDING AND ASSESSMENT OF A TWO-COMPONENT G-PRIOR IN VARIABLE SELECTION

by

Farnaz Solatikia

A Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

Major: Public Health

The University of Memphis
December 2019
Abstract

The selection of variables in regression problems has occupied the minds of many statisticians. Several Bayesian variable selection methods have been developed, and we concentrate on the following method.

This thesis which is a continuation of a project by Zhang et al.2016 published in Bayesian Analysis in 2016, first describes the general idea behind Bayes Inference, various sampling methods based on Bayes theorem. Then we present a Bayesian variable selection method based on an extension of the Zellner’s g-prior in linear models. More specifically, we propose a two-component G-prior, wherein a tuning parameter, calibrated by use of pseudo variables, is introduced to adjust the distance between the two components. We Assess the impact of tuning parameter b, the distance between important and unimportant variables, on the selection of variables by controlling Bayesian false model selection rate with respect to unimportant variables based on creating pseudo variables. We show that implementing the proposed prior in variable selection is more efficient than using the Zellner’s g-prior.
Acknowledgments

I am forever grateful for the continued support, advice, patience, valuable time of my supervisor Dr. Zhang. This thesis would not have been possible without her help, and guidance.

I would like to thank my thesis committee members Dr. Jiang and Dr. Bowman for their time, and valuable comments throughout the preparation and review of this document.

I would like to express my special appreciation and heartfelt thankfulness to my dear Mother and my deceased Father, who passed away this September, for their great role in my life and their numerous sacrifices for me.
Table of contents

1 Introduction 1

2 Linear Regression Model and variable selection 5
   2.1 Parameter estimation in the Bayesian framework 7
   2.2 Bayesian Variable Selection 12
   2.3 Zellner’s $g$-prior in Bayesian variable selection 14

3 Extension of Zellner’s $g$-prior 21
   3.1 The two component $G$-prior 21
   3.2 Prior and hyper-prior distributions for the parameters 23
   3.3 Full conditional posterior distribution and its computation 24
   3.4 Monte Carlo simulation 31
   3.5 Calibration of a tuning parameter 32
   3.6 Properties of the $G$-prior 37

4 Simulation 43
   4.1 Simulation Design 43
   4.2 Results and Discussion 44

5 Conclusion 48

Appendix 50
Chapter 1

Introduction

Regression analysis is a widely applied statistical method to investigate the influence of regressors on a response variable. In the simplest case of a normal linear regression model, it is assumed that the mean of the response variable can be described as a linear function of influential regressors. Selection of the regressors is substantial. If more regressors are included in the model, a higher proportion of the response variability can be explained. On the other hand, overfitting, i.e. including regressors with zero effect worsens the predictive performance of the model and causes loss of efficiency.

An important problem in statistical analysis is the choice of an optimal model from a set of a priori plausible models. For scientists, the regression model is not more than an instrument to represent the relationship between causes and effects of the reality which they want to detect and discover. The inclusion of non relevant quantities in the model or the exclusion of causal factors from the model yields wrong scientific conclusions and interpretations how things work. So correct classification of these two types of regressors is a challenging goal for the statistical analysis. Thus, methods for variable selection are needed to identify zero and non-zero effects.

Commonly used classical variable selection methods are backward, forward and stepwise selection, where in every step regressors are added to the model or eliminated from the
model according to a precisely defined testing schedule. Moreover, the methods built on Akaike information criterion (AIC) and Bayesian information criterion (BIC) (a criterion for model selection among a finite set of models based on the likelihood function) are often used to assess the trade-off between model complexity and goodness-of-fit of the competing models.

Bayesian methods for variable selection are discussed rigorously in the literature as well. Compared to the frequentist approaches for variable selection, one major advantage of Bayesian methods exists in their ability to incorporate prior knowledge into the selection process. In addition, rather than selecting a unique set of variables as in frequentist approaches, Bayesian methods estimate the posterior probabilities for all models under consideration. George and McCulloch\(^{1993}\) proposed an empirical method of stochastic search variable selection (SSVS). Each $\beta_j$ is selected or rejected based on a Monte Carlo average of $\gamma_j$, coming from a Gibbs sampler. Such Monte Carlo average of $\gamma_j$ is called the posterior inclusion probability (PIP) of $\beta_j$. Similar work can be seen in Barbieri and Berger\(^{2004}\), 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, Fokou\(^{2007}\) modified the method in Barbieri and Berger 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. O’Hara and Sillanpaa\(^{2011}\) provides a thorough review of different methods in Bayesian variable selection. They reviewed the developed methods such as Kuo and Mallick, Gibbs Variable Selection (GVS), Stochastic Search Variable Selection (SSVS), adaptive shrinkage with Jeffreys’ prior or a Laplacian prior, and reversible jump MCMC in the context of their different properties. Their results suggested that SSVS, reversible jump MCMC and adaptive shrinkage methods can all work well, but the choice of which method is better will depend on the priors that are used, and also on how they are implemented.

Most work in this area uses g-priors proposed in Zellner\(^{1986}\) and the spike and slab
models (Mitchell and Beauchamp, 2001) for the prior distribution of regression coefficients. Zellner (1986) introduced the g-prior as a reference or default prior for use with Gaussian linear regression models. Recently, variants of the g-prior have been proposed for use with generalized linear models. Choosing either too big or too small g will lead to exclusion of important variables (Lindley, 1957). This Lindley's paradox of Zellner's g-prior in variable selection confesses the importance of proper choice of g (Lindley, Bartlett, 1957). To this end, Liang et al. (2008) proposed using mixtures of g-priors in variable selection. Maruyama and George (2011), in light of an appropriate distribution of prior information on different variables, advocated a generalized g-prior based on the singular value decomposition of the design matrix. However, a single g is a global shrinkage factor that applies to all predictors and will equally shrink each coordinate. This choice is reasonable if we do not know anything about the predictors or if they are equivalent in some sense as noted in a recent study. Som et al. (2014) introduced the block hyper-g/n prior for so-called "poly-shrinkage", which is a collection of ordinary mixtures of g-priors applied separately to groups of predictors. Their motivation is to avoid certain paradoxes, related to different asymptotic behaviour for different subsets of predictors. Min and Sun (2001) considered the situation of grouped covariates (occurring, for example, in ANOVA models where each factor has various levels) and propose separate g-priors for the associated groups of regression coefficients. This also circumvents the fact that in ANOVA models the full design matrix is often not of full rank. In 2016, Zichen Ma introduced a methodology for Bayesian variable selection in linear regression that is independent of the traditional indicator method.

An interesting idea is used by Zhang et al. (2016) where a two-component extension of the g-prior is proposed, with each regressor being assigned one of two possible values for g. In this article, their focus is on the g-prior, a multivariate normal distribution with mean zero and covariance being a proportion g to the covariance of regression coefficient maximum likelihood estimators. Their prior is proper by treating the intercept as part of the regression vector in the g-prior and by using a "vague" proper prior (Note that this implies the necessity
to choose the associated hyper parameters in a sensible manner, which is nontrivial as what
is sensible depends on the scaling of the data) on $\sigma^2$. They focus mostly on variable
selection. They extend the Zellner’s g-prior from a single g to a diagonal matrix G to
incorporate information in the predictors. By injecting a diagonal matrix G to the variance of
the prior, each $g_j$ on the diagonal serves as a variance stabilizer such that the promising
variables are selected based on the $g_j$’s that are close to 0. Then they introduced a tuning
parameter to optimize the prior. This tuning parameter is adaptively selected with the help of
carefully created pseudo-variables. Compared to the existing methods, a major contribution
of this study exists in the extension of the classical g-prior and the inclusion of
pseudo-variables to optimize the extended g-prior.

Their work is motivated by a genetic and epidemiologic study related to the severity of
ideopathic pulmonary fibrosis (IPF). IPF is a chronical lung disease characterized by scarring
of the supporting framework of lungs. The incidence and mortality continue to increase with
more men being afflicted than women. The goal of that study is to identify genes whose
expression levels are associated with the severity of IPF, or genes such that the association is
gender-specific. For this type of applications, variable selection on linear regression models is
commonly used.

In this thesis we try to understand and assess of this new methodology for Bayesian
variable selection in multiple linear regression that is independent of the standard indicator
vector method. Serving as an extension of Zellner’s g-prior, we extend the original scalar g to
a diagonal matrix $G$ that controls the stability of the prior on the coefficients $\beta$. The
formulation of a two-component G-prior, the efficiency of the prior in terms of variable
selection, the prior distributions of other parameters, and the posterior distributions are
discussion. An adaptive method to determine the value of a tuning parameter is also
presented. Finally, we summarize and discuss our findings by simulation which used to
demonstrate and evaluate the proposed method, and its application to the expression data in
thesis.
Chapter 2

Linear Regression Model and variable selection

In statistics regression analysis is a common tool to analyze the relationship between a dependent variable called the response and independent variables called covariates or regressors. It is assumed that the regressors have an effect on the response variable, and thus the researcher wants to quantify this influence. The simplest functional relationship between response variable and potentially influential variables is given by a linear regression model, in which the response can be described as a linear combination of the covariates with appropriate weights called regressor coefficients.

The frequentist view of linear regression is the model assumes that the response variable also called the dependent variable is a linear combination of weights known as the model parameters multiplied by a set of predictor variables. The full formula also includes an error term to account for random sampling noise.

**Goal**: find a proper representation of the distribution, \( f(y|\theta, x) \), of a response or outcome variable \( y \) given an explanatory variables \( x \), based on a sample of \( (x,y)_i \)’s.

The multivariate normal regression model is a distribution \( Y|\beta, \sigma^2, X \sim MNV(X\beta, \sigma^2 I_n) \) where \( I \) is the \( p \times p \) identity matrix, \( Y \) a \( n \times 1 \) vector observed dependent variable, \( X \) a \( n \times p \)
matrix of explanatory variables, $\beta$ a $p \times 1$ vector of unknown regression coefficients, and $\sigma$ is an unknown positive scalar namely

$$Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$  \hspace{1cm} (1)

and $V(y_i | \sigma^2, X) = \sigma^2$. This completely specifies the probability density of the data:

$$p(y_1, \ldots, y_n | x_1, \ldots, x_n, \beta, \sigma^2) = \prod_{i=1}^{n} p(y_i | x_i, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 \right\}$$  \hspace{1cm} (3)

So the likelihood of ordinary normal linear model

$$
\ell(\beta, \sigma^2 | y, X) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \right]
$$  \hspace{1cm} (4)

is big when sum of squared residuals (SSR) is small.

$$SSR(\beta) = \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 = (y - X\beta)^T (y - X\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta.$$  \hspace{1cm} (5)

What value of $\beta$ makes this the smallest?

The MLE of $\beta$ is solution of the least squares minimisation problem,

$$\min_{\beta} \text{SSR}(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2$$

$$\frac{d}{d\beta} \text{SSR}(\beta) = -2X^T Y + 2X^T X \beta = 0.$$  \hspace{1cm} (6)
Usually the unknown coefficient vector is estimated by the ordinary-least-square method (OLS), 
\[ \hat{\beta} = (X^T X)^{-1} X^T Y \] 
of \( \beta \) and this value makes \( SSR \) the smallest.

The goal of learning a linear model from training data is to find the coefficients that best explain the data. In frequentist linear regression, the best explanation is taken to mean the coefficients that minimize the sum of squared residuals (SSR). The method of fitting the model parameters by minimizing the RSS is called Ordinary Least Squares (OLS).

- \( \hat{\beta} \) is an unbiased estimator of \( \beta \),
- \( V(\hat{\beta}|\sigma^2, X) = \sigma^2 (X^T X)^{-1} \),
- \( \hat{\beta} \) is the best linear unbiased estimator of \( \beta \): for all \( a \in \mathbb{R}^{p+1} \), and for any unbiased linear estimator \( \tilde{\beta} \) of \( \beta \), \( V(a^T \tilde{\beta}|\sigma^2, X) \leq V(a^T \hat{\beta}|\sigma^2, X) \),
- Unbiased estimator of \( \sigma^2 \), \( \hat{\sigma}^2 = \frac{1}{n-p-1} (y - X \hat{\beta})^T (y - X \hat{\beta}) = \frac{s^2}{n-p-1} \),
- If the Jefferys prior is flat, \( \pi(\beta) \propto 1 \) then the posterior is \( \beta \sim N(\hat{\beta}_{OLS}, \Sigma_{OLS}) \) where \( \hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \) and \( \Sigma_{OLS} = \sigma^2 (X^T X)^{-1} \).

The posterior is proper if and only if \( X^T X \) is nonsingular and with a flat prior, estimates and inference are identical to least squares.

In quite a few areas where the linear model applies, an interesting yet very important fact is that only a small portion of variables affect the response whereas others are trivial (Jeffreys and Berger, 1992). A great many authors have discussed this topic from both the frequentist (for example, Ullah and Wang, 2013) and the Bayesian perspective (Walli and Wagner, 2011). In this proposal, we proceed following the Bayesian path.

### 2.1 Parameter estimation in the Bayesian framework

Bayes’ original theorem applied to point probabilities. The basic theorem states simply:

\[ p(B|A) = \frac{p(A|B)p(B)}{p(A)}. \] (7)
Bayes’ theorem, and indeed, its repeated application in cases is beyond mathematical dispute. However, Bayesian statistics typically involves using probability distributions rather than point probabilities for the quantities in the theorem. This process of repeating the test and recomputing the probability of interest is the basic process of concern in Bayesian statistics. From a Bayesian perspective, we begin with some prior probability for some event, and we update this prior probability with new information to obtain a posterior probability. The posterior probability can then be used as a prior probability in a subsequent analysis. From a Bayesian point of view, this is an appropriate strategy for conducting scientific research: We continue to gather data to evaluate a particular scientific hypothesis; we do not begin anew (ignorant) each time we attempt to answer a hypothesis, because previous research provides us with a priori information concerning the merit of the hypothesis.

Put generally, the goal of Bayesian statistics is to represent prior uncertainty about model parameters with a probability distribution and to update this prior uncertainty with current data to produce a posterior probability distribution for the parameter that contains less uncertainty. This perspective implies a subjective view of probability—probability represents uncertainty—and it contrasts with the classical perspective. From the Bayesian perspective, any quantity for which the true value is uncertain, including model parameters, can be represented with probability distributions. From the classical perspective, however, it is unacceptable to place probability distributions on parameters, because parameters are assumed to be fixed quantities: Only the data are random, and thus, probability distributions can only be used to represent the data.

Bayes’ Theorem, expressed in terms of probability distributions, appears as:

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)},$$  \hspace{1cm} (8)

where $f(\theta|data)$ is the posterior distribution for the parameter $\theta$, $f(data|\theta)$ is the sampling density for the data—which is proportional to the Likelihood function, only differing by a
constant that makes it a proper density function. \( f(\theta) \) is the prior distribution for the parameter, and \( f(data) \) or the normalizing constant, which is simply the integral of numerator over the parameter space (which may be finite- or infinite-dimensional), is called the marginal probability of the data. For a continuous sample space, this marginal probability is computed as: 

\[
f(data) = \int f(data | \theta) f(\theta) d\theta,
\]

the integral of the sampling density multiplied by the prior over the sample space for \( \theta \). This quantity is sometimes called the "marginal likelihood" for the data and acts as a normalizing constant to make the posterior density proper. Because this denominator simply scales the posterior density to make it a proper density, and because the sampling density is proportional to the likelihood function, Bayes’ Theorem for probability distributions is often stated as:

\[
Posterior \propto Likelihood \times Prior. \tag{9}
\]

In the Bayesian viewpoint, we formulate linear regression using probability distributions rather than point estimates. The response is not estimated as a single value, but is assumed to be drawn from a probability distribution. The aim of Bayesian Linear Regression is not to find the single “best” value of the model parameters, but rather to determine the posterior distribution for the model parameters. Not only is the response generated from a probability distribution, but the model parameters are assumed to come from a distribution as well. In contrast to OLS, we have a posterior distribution for the model parameters that is proportional to the likelihood of the data multiplied by the prior probability of the parameters. Here we can observe the two primary benefits of Bayesian Linear Regression.

- Priors: If we have domain knowledge, or a guess for what the model parameters should be, we can include them in our model, unlike in the frequentist approach which assumes everything there is to know about the parameters comes from the data. If we don’t have any estimates ahead of time, we can use non-informative priors for the parameters such as a normal distribution. Often it is possible to utilize the fact that the
model has a set of parameters which have a similar role. It is, for instance, reasonable to assume that all biases in an MLP network have a similar distribution. This knowledge can be utilised by modelling the distribution by a common parametrised distribution. Then the prior needs to be determined to these common parameters, called hyperparameters but as they control the distribution of a set of parameters.

- Posterior: The result of performing Bayesian Linear Regression is a distribution of possible model parameters based on the data and the prior. This allows us to quantify our uncertainty about the model: if we have fewer data points, the posterior distribution will be more spread out.

The Bayesian model starts with the same model as the classical frequentist approach:

\[ Y_i = \beta x_i + e_i, \quad i = 1, \ldots, n \]

with the assumption that the errors are independent and identically distributed as normal random variables with mean zero and constant variance \( \sigma^2 \). This assumption is exactly the same as in the classical inference case for testing and constructing confidence intervals for \( \beta \) and \( \sigma^2 \). Our goal is to update the distributions of the unknown parameters \( \beta, \sigma^2 \), based on the data \( x_1, y_1, \ldots, x_n, y_n \), where \( n \) is the number of observations. Under the assumption that the errors are normally distributed with constant variance \( \sigma^2 \), we have for the random variable of each response \( Y_i \), conditioning on the observed data \( x_i \) and the parameters \( \beta, \sigma^2 \), is normally distributed:

\[
y_i|x_i, \beta, \sigma^2 \sim N(\beta x_i, \sigma^2), \quad i = 1, \ldots, n.
\]

That is, the likelihood of \( Y \) given \( X \) which is the product of each likelihood \( p(y_i|x_i, \beta, \sigma^2) \), since we assume each response \( y_i \) is independent from each other.

A Bayesian analysis also requires priors for \( \beta \) and \( \sigma^2 \). We will focus on prior specification since this piece is uniquely Bayesian. Many priors for \( \beta \) have been considered: The Jeffreys’ prior is improper prior and flat \( p(\beta) = 1 \) but the posterior is proper under the same conditions required by least squares. If \( \sigma^2 \) is known then \( \beta|Y \sim N(\hat{\beta}_{OLS}, \sigma^2(X^TX)^{-1}) \). Therefore, the
results should be similar to least squares. Of course we rarely know $\sigma^2$, typically the error variance follows an $\text{Inv-Gamma}(a,b)$ prior with $a$ and $b$ set to be small, say $a = b = 0.01$. In this case the posterior of $\beta$ follows a multivariate $t$ centered on $\hat{\beta}_{\text{OLS}}$, again, the results are similar to OLS. The objective Bayes Jeffreys prior for $\theta = (\beta, \sigma^2)$ is $p(\beta, \sigma^2) = 1/\sigma^2$ which is the limit as $a, b \to 0$.

In the Bayesian approach probability distributions are used to quantify uncertainty. Thus, in contrast to the frequentist approach, a joint stochastic model for response and parameters $(\beta, \sigma^2)$ is specified. The distribution of the dependent variable $y$ is specified conditional on the parameters $\beta$, $\sigma^2$ and the sampling density of the data $p(y_1, \ldots, y_n|x_1, \ldots, x_n, \beta, \sigma^2)$, as a function of $\beta$. We first consider the case under the reference prior, which is our standard non-informative prior. Using the reference prior, we will obtain familiar distributions as the posterior distributions of $\beta$, and $\sigma^2$, which gives the analogue to the frequentist results. Then we apply the Bayes' rule to derive the joint posterior distribution after observing data $y_1, \ldots, y_n$. Bayes' rule states that the joint posterior distribution of $\beta$, and $\sigma^2$ is proportional to the product of the likelihood and the joint prior distribution:

$$p(\beta, \sigma^2|Y, X) \propto L(Y, X)p(\beta, \sigma^2).$$  \hspace{1cm} (11)

To obtain the marginal posterior distribution of $\beta$, we need to integrate $\sigma^2$ out from the joint posterior distribution. The role that $\beta$ plays in the exponent looks very similar to that played by $Y$, and the distribution of $Y$ is multivariate normal. In order to estimate the unknown quantities in the Bayesian models, Markov Chain Monte Carlo (MCMC) methods are useful tools to sample from those posterior distributions that have no closed form. The Gibbs Sampling is a stochastic simulation algorithm via Markov Chain utilized when the joint posterior distribution has no closed form but all its conditional distributions do. It is one way (among many) to estimate a Bayesian model.
2.2 Bayesian Variable Selection

An important problem in linear regression is variable selection. The aim of this procedure is to reduce the whole set of predictors to a best subset. Identifying the predictors that significantly affect the response variable is crucial; it is also necessary since the redundant covariates need be removed so that the model can be precise, simple and can provide accurate predictions. To be more specific, the problem of collinearity can arise if two or more predictors are explaining the same thing. In other words, in terms of likelihood, we would expect that the likelihood value will increase as the number of predictors goes up, i.e., the complexity of the model becomes bigger. The problem is that a model with more predictor variables will always do better than the simpler model. But this leads to a problem called “overfitting”, which will destroy the prediction accuracy. Hence, a good variable selection strategy is as crucial as the problem of variable selection itself.

It is important to be able to decide which variables—within a large pool of potential explanatory variables—should be kept in a model that balances good explanatory power with good estimation performance. In many analyses, this reduces to the choice of which subset of variables should be included into the model. This problem has exercised the minds of many statisticians, leading to a variety of algorithms for searching the model space and selection criteria for choosing between competing models. In the Bayesian framework, the model selection problem is transformed to the form of parameter estimation: rather than searching for the single optimal model, a Bayesian will attempt to estimate the posterior probability of all models within the considered class of models. In many cases, this question is asked in variable-specific form: i.e. the task is to estimate the marginal posterior probability that a variable should be in the model. A Bayesian analysis starts by assigning prior distributions to the unknown model parameters $\beta$. However, usually there is not only uncertainty with respect to the model parameters, but also to the model itself.

We consider a dependent random variable $y$ and a set of $p$ potential explanatory variables. Let $\gamma$ be the model index contained in some model space $\Gamma$. Typically, the variable
selection problem is considered, where \( \vec{\gamma} \in \{0, 1\}^p \) collects binary inclusion indicators for all \( p \) available covariates. At this stage, we assume that every subset of \( q \) explanatory variables could make a proper set of explanatory variables for the regression of \( y \). The only restriction we impose is that the intercept (that is, the constant variable) is included in every model. There are thus \( 2^p \) models in competition and we are looking for a procedure that selects the “best” model, that is, the “most relevant” explanatory variables. The variable-selection procedure can be seen as a two-stage estimation setting first estimate the indicator of the model (within the collection of models), then estimate the parameters corresponding to this very model. Each of the \( 2^p \) models \( M_\gamma \) associated with binary indicator vector \( \vec{\gamma} \in \Gamma = \{0, 1\}^p \) where \( \gamma_i = 1 \) means that the variable \( x_i \) is included in the model \( M_\gamma \) and \( \gamma_i = 0 \) means that excluded from the model. The number of variables included in the model \( M_\gamma \) is \( \Lambda_\gamma = 1^T_n \gamma \). We also define \( \beta_\gamma \) sub-vector of \( \beta \) and \( X_\gamma \) as the sub-matrix of \( X \) where only the column \( x_j \) is included in the model \( M_\gamma \) have been left. The sub-model \( M_\gamma \) is thus defined as

\[
Y = X_\gamma \beta_\gamma + e, \quad e \sim MVN(0, \sigma^2 I)
\]

\[
Y | \gamma, \beta_\gamma, \sigma^2, X \sim MVN(X_\gamma \beta_\gamma, \sigma^2 I_n)
\]

where \( \beta_\gamma \) and \( \sigma^2 \) are the unknown parameters, \( \sigma^2 \) is common to all models and thus uses the same prior for all models. We cannot expect a practitioner to specify a prior on every \( M_\gamma \) (\( 2^k \) models) in a completely subjective and autonomous manner. We derive all priors from a single global prior associated with the so-called full model that corresponds to \( \vec{\gamma} = (1, \ldots, 1) \).

**Prior Distributions for Model Choice:**

Although there are many possible ways of defining the prior on the model index \( \gamma \), we opt for the uniform prior \( \pi(\gamma) = 2^{-p} \). The posterior distribution of \( \gamma \) is central to the variable-selection methodology since it is proportional to the marginal density of \( y \) in \( M_\gamma \). In addition, for prediction purposes, the prediction distribution can be obtained by averaging over all models, the weights being the model probabilities. When the number \( p \) of variables is
large, it becomes impossible to compute the posterior probabilities for the whole series of $2^p$
models. Assume we know the marginal posterior function of $\gamma$ then need a tailored algorithm
that samples from $\pi(\gamma|Y, X)$ and thus selects the most likely models, without computing first
all the values of $\pi(\gamma|Y, X)$. This can be done rather naturally by Gibbs sampling, given the
availability of the full conditional posterior probabilities of the $\gamma^*_j$’s. If

$$\gamma^*_{j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$$

the full conditional distribution $\pi(\gamma_j|Y, \gamma^*_{j}, X)$ of $\gamma_j$ is
proportional to $\pi(\gamma|Y, X)$ and can be computed in both $\gamma_j = 0$ and $\gamma_j = 1$ at no cost.

Algorithm Gibbs Sampler for Variable Selection

| Initialization: Draw $\gamma^0$ from the uniform
distribution on $\Gamma$. | Iteration $t$: Given $(\gamma_1^{(t-1)}, \ldots, \gamma_p^{(t-1)})$, generate |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $\gamma_1^{(t)}$ according to $\pi(\gamma_1</td>
<td>Y, \gamma_2^{(t-1)}, \ldots, \gamma_p^{(t-1)}, X)$</td>
</tr>
<tr>
<td>2. $\gamma_2^{(t)}$ according to $\pi(\gamma_2</td>
<td>Y, \gamma_1^{(t)}, \gamma_3^{(t-1)}, \ldots, \gamma_p^{(t-1)}, X)$</td>
</tr>
<tr>
<td>\vdots</td>
<td></td>
</tr>
<tr>
<td>p. $\gamma_p^{(t)}$ according to $\pi(\gamma_p</td>
<td>Y, \gamma_1^{(t)}, \ldots, \gamma_{p-1}^{(t)}, X)$</td>
</tr>
</tbody>
</table>

After a large number of iterations of this algorithm that is, when the sampler is supposed to
have converged or more accurately when the sampler has sufficiently explored the support of
the target distribution, its output can be used to approximate the posterior probabilities
$\pi(\gamma|Y, X)$ by empirical averages based on the Gibbs output.

### 2.3 Zellner’s $g$-prior in Bayesian variable selection

In the Bayesian stochastic search variable selection framework, a common prior distribution
for the regression coefficients is the g-prior of Zellner (1986). He in his paper mentioned that
assessing the informative prior distribution for the coefficient parameters is important. He
proposed a reference informative prior called “g-prior” which is easy to evaluate the prior
covariance for the elements of $\beta$. However, there are two standard cases in which the
associated covariance matrix does not exist, and the conventional prior of Zellner can not be used: if the number of observations is lower than the number of variables (large p and small n paradigm), or if some variables are linear combinations of others. In such situations a prior distribution derived from the prior of Zellner can be used, by introducing a ridge parameter.

Suppose the prior and posterior distribution in Bayesian inference for regression models be

\[
prior \quad \beta \sim \text{MVN}(\beta_0, \Sigma_0)
\]

\[
sampling \ model \quad Y \sim \text{MVN}(X\beta, \sigma^2 I)
\]

\[
posterior \quad \beta | Y, X \sim \text{MVN}(\beta_n, \Sigma_n)
\]

This suggests the likelihood function and a multivariate normal prior distribution for \(\beta\) as following

\[
p(Y|X, \beta, \sigma^2) \propto \exp\left\{ -\frac{1}{\sigma^2} \text{SSR}(\beta) \right\}
\]

\[
= \exp\left\{ -\frac{1}{\sigma^2} [Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta] \right\}
\]

\[
p(\beta|Y, X, \sigma^2) \propto p(Y|X, \beta, \sigma^2) \times p(\beta)
\]

\[
\propto \exp\left\{ -\frac{1}{2} \left( -2\beta^T X^T Y / \sigma^2 + \beta^T X^T X \beta / \sigma^2 \right) - \frac{1}{2} \left( -2\beta^T \Sigma_0^{-1} \beta_0 + \beta^T \Sigma_0^{-1} \beta \right) \right\}
\]

\[
= \exp\left\{ \beta^T (\Sigma_0^{-1} \beta_0 + X^T Y / \sigma^2) - \frac{1}{2} \beta^T (\Sigma_0^{-1} + X^T X / \sigma^2) \beta \right\}
\]

We recognize this as being proportional to a multivariate normal density, with

\[
\Sigma_n = \text{Var}(\beta|Y, X, \sigma^2) = (\Sigma_0^{-1} + X^T X / \sigma^2)^{-1}
\]

\[
\beta_n = \text{E}(\beta|Y, X, \sigma^2) = (\Sigma_0^{-1} + X^T X / \sigma^2)^{-1} (\Sigma_0^{-1} \beta_0 + X^T Y / \sigma^2)
\]

If the elements of the prior precision matrix \(\Sigma_0^{-1}\) are small in magnitude, then the conditional expectation \(\text{E}(\beta|Y, X, \sigma^2)\) is approximately equal to \((X^T X)^{-1} X^T Y\), the least squares estimate.
On the other hand, if the measurement precision is very small ($\sigma^2$ is very large), then the expectation is approximately $\beta_0$, the prior expectation.

Notice that if $\Sigma_0^{-1} << X^T X / \sigma^2$, then $\beta_n \sim \hat{\beta}_{ols}$ and also if $\Sigma_0^{-1} >> X^T X / \sigma^2$, then $\beta_n \sim \hat{\beta}_0$.

Now how can pick $\beta_0$, and $\Sigma_0$?

In constructing a family of priors for a Gaussian regression model $Y = X\beta + e$, Zellner 1986 suggested a particular form of the conjugate Normal–Gamma family, namely, a g prior, $\beta \sim N(\beta_0, g\sigma^2(X^T X)^{-1})$ and $p(\sigma^2) \propto 1/\sigma^2$, where the prior mean $\beta_0$ is taken as the anticipated value of $\beta$ based on imaginary data and the prior covariance matrix of $\beta$ is a scalar multiple $g$ of the Fisher information matrix, which depends on the observed data through the design matrix $X$. This, along with arguments based on orthogonal parameterizations and invariance to scale and location transformations (Jeffreys, Berger 1992), has led to the adoption of $\beta_0 = 0$, this prior is proper assuming $X$ is full rank and the posterior mean is $\frac{1}{g+1}\hat{\beta}_{OLS}$. This shrinks the least estimates towards zero and $g$ controls the amount of shrinkage. In the Bayesian setting the coefficient $\beta$ is usually given a conventional g-prior $N(0, g\sigma^2(X^T X)^{-1})$, introduced in Zellner 1986. The g-prior has been given much attention in Bayesian variable selection primarily because it leads to a computationally tractable Bayes Factor. By introducing an indicator vector, variables are selected and different subsets of variables are compared to each other, or to a reference, based on the value of Bayes Factor. Multiple works have been done to review this methodology. For a recent one, see Dey and Fokou 2007.

The variance of the OLS estimate $\hat{\beta}_{ols}$ is

$$Var[\hat{\beta}_{ols}] = \sigma^2(X^T X)^{-1} = \frac{\sigma^2}{n} (X^T X / n)^{-1}$$

$$Var[\beta]_{g prior} = g\sigma^2(X^T X)^{-1} = \frac{\sigma^2}{n/g} (X^T X / n)^{-1}$$

This is roughly the uncertainty in $\beta$ from n observations. For example, $g = n$ means the prior has the same amount of info as 1 obs.
The posterior distributions under the g-prior is \( \beta | y, X, \sigma^2 \sim \text{MVN}(\beta_n, \Sigma_n) \) where

\[
\Sigma_n = \text{Var}[\beta | y, X, \sigma^2] = \frac{g}{g+1} \sigma^2 (X^T X)^{-1}
\]

\[
\beta_n = E[\beta | y, X, \sigma^2] = \frac{g}{g+1} (X^T X)^{-1} X^T y
\]  

(18)

Notes that the posterior mean estimate \( \hat{\beta}_n \) is simply \( \frac{g}{g+1} \hat{\beta}_{ols} \) and the posterior variance of \( \beta \) is simply \( \frac{g}{g+1} \text{Var}[\hat{\beta}_{ols}] \). \( g \) shrinks the coefficients and can prevent over-fitting to the data, and if \( g = n \), then as \( n \) increases, inference approximates that using \( \hat{\beta}_{ols} \).

As in most normal sampling problems, the semi-conjugate prior distribution for \( \sigma^2 \) is an inverse-gamma distribution. If \( \sigma^2 \sim \text{Inv-Gamma}(v_0/2, v_0 \sigma_0^2/2) \), then

\[
p(\sigma^2 | y, X, \beta) \propto p(\sigma^2) p(y | X, \beta, \frac{1}{\sigma^2})
\]

\[
\propto (\frac{1}{\sigma^2})^{v_0/2-1} \exp(-\frac{1}{\sigma^2} \times v_0 \sigma_0^2/2) \times (\frac{1}{\sigma^2})^{n/2} \exp(-\frac{1}{\sigma^2} \times \text{SSR}(\beta)/2)
\]

\[
= (\frac{1}{\sigma^2})^{(v_0+n)/2-1} \exp(-\frac{1}{\sigma^2} [v_0 \sigma_0^2 + \text{SSR}(\beta)]/2)
\]

is conjugate. The model for the error variance is

prior \( \sigma^2 \sim \text{Inv-Gamma}(v_0/2, v_0 \sigma_0^2/2) \)

sampling model \( Y \sim \text{MVN}(X\beta, \sigma^2 I) \)

(20)

posterior \( \sigma^2 | y, X \sim \text{Inv-Gamma}((v_0+n)/2, (v_0 \sigma_0^2 + \text{SSR})/2) \)

Simulating the joint posterior distribution:

joint distribution \( p(\beta, \sigma^2 | y, X) = p(\beta | y, X, \sigma^2) \times p(\sigma^2 | y, X) \)

(21)

simulation \( \{\beta, \sigma^2\} \sim p(\beta, \sigma^2 | y, X) \iff \sigma^2 \sim p(\sigma^2 | y, X), \beta \sim p(\beta | y, X, \sigma^2) \)

To simulate \( \{\beta, \sigma^2\} \) i.e. to implement the Gibbs sampler in the linear regression model, we can proceed as follows. First simulate \( \sigma^2 \) from \( p(\sigma^2 | y, X) \) then use this \( \sigma^2 \) to simulate \( \beta \) from \( p(\beta | y, X, \sigma^2) \). Repeat 1000’s of times to obtain MC samples: \( \{\beta, \sigma^2\}_1, ..., \{\beta, \sigma^2\}_S \).
The g-prior can roughly be viewed as the uncertainty from \( n/g \) observations. This modeling somehow appears as a data-dependent prior through its dependence on \( X \) is decomposed as a (conditional) Gaussian prior for \( \beta \) and an improper (Jeffreys) prior for \((\alpha, \sigma^2)\). The experimenter thus restricts prior determination to the choices of \( \tilde{\beta} \) and of the constant \( g \) namely, \( \beta|\sigma^2, X \sim \text{MVN}(\tilde{\beta}, g\sigma^2(X^TX)^{-1}) \) and a non-informative prior distribution is imposed on the pair \((\alpha, \sigma^2)\), \( \pi(\sigma^2|X) \propto \sigma^{-2} \). With this prior model, the posterior simplifies into

\[
\pi(\beta, \sigma^2|y, X) \propto f(y|\beta, \sigma^2, X) \pi(\beta, \sigma^2|X)
\]

\[
\propto (\sigma^2)^{-\left(n/2 + 1 + k/2\right)} \exp\left[\frac{1}{2\sigma^2} (y - X\hat{\beta})^T (y - X\hat{\beta}) - \frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^TX (\beta - \hat{\beta}) - \frac{1}{2g\sigma^2} (\beta - \tilde{\beta})^T X^TX (\beta - \tilde{\beta})\right],
\]

because \( X^TX \) used in both prior and likelihood therefore,

\[
\beta|\sigma^2, y, X \sim \text{N}_{p+1}\left(\frac{g}{g+1}(\tilde{\beta}/g + \hat{\beta}), \frac{\sigma^2 g}{g+1} (X^TX)^{-1}\right)
\]

\[
\sigma^2|y, X \sim \text{Inv.Gam}\left(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(g+1)}(\tilde{\beta} - \hat{\beta})^T X^TX (\tilde{\beta} - \hat{\beta})\right)
\]

\[
\beta|y, X \sim T_{k+1}\left(n, \frac{g}{g+1} \left(\frac{\tilde{\beta}}{g} + \hat{\beta}\right), \frac{g(s^2 + (\tilde{\beta} - \hat{\beta})^T X^TX (\tilde{\beta} - \hat{\beta})/(g+1))}{n(g+1)} (X^TX)^{-1}\right).
\]

The Bayes estimators of \( \beta \) and \( \sigma^2 \) are given by

\[
E^\pi[\beta|y, X] = \frac{1}{g+1} (\tilde{\beta} + g\hat{\beta})
\]

and

\[
E^\pi[\sigma^2|y, X] = \frac{s^2 + (\tilde{\beta} - \hat{\beta})^T X^TX (\tilde{\beta} - \hat{\beta})/(g+1)}{n-2}.
\]

Only when \( g \) goes to infinity does the influence of the prior vanish!
Moreover,

$$V^\pi[\beta | y, X] = \frac{c(s^2 + (\tilde{\beta} - \hat{\beta})^T X^T X (\tilde{\beta} - \hat{\beta})/(c+1))}{n(c+1)} (X^T X)^{-1}. \quad (25)$$

$g$ can be interpreted as a measure of the amount of information available in the prior relative to the sample, for instance, setting $1/g = 0.5$ gives the prior the same weight as 50% of the sample.

1. For the full model, Zellner’s $g$-prior:

$$\beta | \sigma^2, X \sim N_{p+1} (\tilde{\beta}, g \sigma^2 (X^T X)^{-1}) \quad \text{and} \quad \sigma^2 \sim \pi(\sigma^2 | X) = \sigma^{-2} \quad (26)$$

2. For each sub-model $M_\gamma$, the prior distribution of $\beta_\gamma$ conditional on $\sigma^2$ is fixed as

$$\beta_\gamma | \gamma, \sigma^2 \sim N_{q+1} (\tilde{\beta}_\gamma, g \sigma^2 (X^T_{\gamma} X_{\gamma})^{-1}), \quad (27)$$

where $\tilde{\beta}_\gamma = (X^T_{\gamma} X_{\gamma})^{-1} X^T_{\gamma} \tilde{\beta}$ and same prior on $\sigma^2$.

The joint prior for model $M_\gamma$ is the improper prior

$$\pi(\beta_\gamma, \sigma^2 | \gamma) \propto (\sigma^2)^{-(q_\gamma + 1)/2 - 1} \exp \left[ - \frac{1}{2g\sigma^2} (\beta_\gamma - \tilde{\beta}_\gamma)^T (X^T_{\gamma} X_{\gamma}) (\beta_\gamma - \tilde{\beta}_\gamma) \right] \quad (28)$$

We need to determine how to set $g$ (Convenient tool for translating prior information on $\beta$). Zellner’s informative $g$ prior intuitively determines how much the prior distribution of $\beta$ contributes to the posterior. For instance, if $g = 0$, the posterior mean fully shrinkages to the prior mean; if $g = 1$, the posterior mean shrinkages 50 percent to the prior mean. In this case, the Bayesian estimate of $\beta$ is the average between the least square estimator and the prior expectation i.e. $E^T(\beta | y, X) = (\tilde{\beta} + \hat{\beta})/2$. If $g$ goes to $\infty$, the prior is a diffuse prior and the posterior mean converges to $\hat{\beta}$. The larger $g$ is, the weaker the prior information and the closer the Bayesian estimator is to the least squares estimator.
Zellner’s g prior is popular in variable selection. It provides a closed form for the marginal likelihood and an explicit expression for Bayes factor, which ensures a fast computation. However, the choice of g is problematic. Zellner himself mentioned that g can depend on the sample size n, e.g., $g \propto 1/n$ or put a prior on g, and g can be integrated out. g could be chosen by finding the maximum posterior probability and corresponding to some popular selection criteria mentioned below. George and Foster\textsuperscript{1993} proposed the empirical Bayes selection criteria which have dimensionality penalties depending on the data. Liang\textsuperscript{2008} proposed Hyper-g priors, which provides robustness to misspecification of g while maintaining the computational efficiency.

Table 2.1: Zellner’s g-prior.

<table>
<thead>
<tr>
<th>Proposal</th>
<th>Reference</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g=n</td>
<td>zellner 1986</td>
<td>Unit Information Prior (UIP)</td>
</tr>
<tr>
<td>g=p(^2)</td>
<td>Foster/Goerge 1994</td>
<td>Risk Information Criterion prior (RIC)</td>
</tr>
<tr>
<td>g=max{n, p(^2}}</td>
<td>Fernandez 2001</td>
<td>Benchmark prior (BRIC)</td>
</tr>
<tr>
<td>g=log(n)</td>
<td>Fernandez 2001</td>
<td>Hannan-Quinn (HQ)</td>
</tr>
<tr>
<td>g=(\hat{g}^\gamma)</td>
<td>Liang 2008</td>
<td>Local Empirical Bayes (LEB)</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g \sim Inv.Gam\left(\frac{1}{2}, \frac{n}{2}\right)</td>
<td>Jeffrey 1961</td>
<td>cauchy prior</td>
</tr>
<tr>
<td>g</td>
<td>a \sim \pi(g) \propto (1 + g)^{-a/2}</td>
<td>Liang 2008</td>
</tr>
<tr>
<td>g</td>
<td>a \sim \pi(g) \propto (1 + g/n)^{-a/2}</td>
<td>Liang 2008</td>
</tr>
</tbody>
</table>
Chapter 3

Extension of Zellner’s g-prior

In this chapter, we present a two-component G method for variable selection, and the formulation of a two-component G-prior, the efficiency of the prior in terms of variable selection, the prior distributions of other parameters, and the posterior distributions are discussion. Also an adaptive method to determine the value of a tuning parameter is presented.

3.1 The two component G-prior

Zellner’s informative g prior intuitively determines how much the prior distribution of $\beta$ contributes to the posterior. He mentioned that g can depend on the sample size $n$ or put a prior on g, and g can be integrated out. We develop an extension of the classical Zellner’s g-prior to a two component G-prior and construct the prior distribution of $\beta$ in two steps.

1. Extension of Zellner’s g-prior:

   We define $g_l = bf_1(n) > 0, g_s = bf_2(n) > 0$ with $f_1(n) = O(n), f_2(n) = O(n^\psi)$,

   $\frac{1}{2} < \psi < 1$, (Big-O notation describes the upper bound, and it means that for large $n$,

   it is less than the function $f(n) = k \times n$ for some fixed $k$ or means the algorithm takes no

   longer than (i.e. $\leq$) that long (as the input size tends to infinity, with relevant constant
factors considered), implying \( f_2(n) = o(f_1(n)) \) as \( n \to \infty \) roughly means \( f_2(n) < f_1(n) \) then \( g_s, g_l \to \infty \), while \( g_s / g_l \to 0 \). The \( b > 1 \) is a tuning parameter and determines the distance between \( g_l \) and \( g_s \) i.e. \( |g_l - g_s| = |b(f_1(n) - f_2(n))| \), and \( g \) is a mixture of two point masses where \( I(.) \) is an indicator variable and \( q \) is the probability of \( g = g_l \) i.e.

\[
p(g) = q I\{g = g_l\} + (1 - q) I\{g = g_s\}. \tag{29}
\]

For a given \( f_1(n) \) and \( f_2(n) \), \( b \) determines the distance between \( g_l \) and \( g_s \), and will be adaptively calibrated via the controlling of false model selection rates using pseudo variables. Thus the ultimate functionality of \( b \) is to optimize \( g_k \) to eliminate unimportant variables. For instance, suppose \( f_1(n) = n, f_2(n) = n^{0.55} \) so \( g_l = bn, g_s = bn^{0.55} \). Next, instead of using one \( g \) to shrink all, we extend the application of \( p(g) \) to each coordinate and define \( G \) as

\[
G = \text{diag}(g_1, ..., g_p), \tag{30}
\]

a diagonal matrix such that each \( g_k \) is chosen for all \( k = 1, ..., p \) according to

\[
p(g_k) = q_k I\{g_k = g_l\} + (1 - q_k) I\{g_k = g_s\}. \tag{31}
\]

Let \( c = (c_1, ..., c_p) \) be a vector of 1s and 0s denoting the choice between \( g_l \) and \( g_s \), respectively. Thus we have

\[
g_k = g_l c_k + g_s (1 - c_k),
\]

with \( p(c_k = 1) = q_k \) for \( k = 1, ..., p \).

II. The prior distribution of \( \beta \) for two component G-prior:

With inject a prior to the coefficient \( \beta \) instead of the covariance matrix of the \( g \)-prior, \( g \sigma^2 (X^T X)^{-1} \), it becomes \( \sigma^2 [(X_\gamma G_\gamma^{-1})^T (X_\gamma G_\gamma^{-1})^{-1}] \) i.e.

\[
\beta_\gamma | X_\gamma, \gamma, \sigma^2, G \sim \mathcal{N} \left( 0, \sigma^2 G_\gamma \left( X_\gamma^T X_\gamma \right)^{-1} G_\gamma \right) \tag{32}
\]
where $\gamma = (\gamma_1, ..., \gamma_p)$, $X^T\gamma X\gamma$ is the prior correlation matrix, and

$$G_\gamma = \text{diag}[g_1, ..., g_p] \text{ and } g_k = g_l c_k + g_s (1 - c_k)$$

(33)

which determines the scaling of the prior covariance matrix. If $g_k = g_s$ for all $k = 1, ..., p$, then the prior is essentially the Zellner’s g-prior, $N(0, g_s^2 (X^T\gamma X\gamma)^{-1} \sigma^2)$; so is the situation if $g_k = g_l$ for all k’s.

### 3.2 Prior and hyper-prior distributions for the parameters

Let $Y \in \mathbb{R}^{n \times 1}$ be an $n$-dimensional row vector; $X \in \mathbb{R}^{n \times p}$ be a $n \times p$-dimensional matrix; and $\beta \in \mathbb{R}^{n \times p}$ be a $p$-dimensional row vector. With this notation, the residual sum of squares, which was $\sum_{i=1}^n (Y_i - \beta X_i)^2$ previously, becomes $(Y - X\beta)^T (Y - X\beta)$. Similarly, where we previously had $\beta^2$ we now have $\beta^T \beta$.

In Bayesian statistics, a hyperprior is a prior distribution on a hyperparameter, that is, on a parameter of a prior distribution. We use independent priors and hyperprior over each predictor:

1. $\beta_\gamma \sim \mathcal{N}(0, \sigma^2 G_\gamma (X^T_\gamma X_\gamma)^{-1} G_\gamma)$ with $\vec{\gamma} = (\gamma_1, ..., \gamma_p)$, and $G_\gamma = \text{diag}[g_1, ..., g_p]$,

2. $\gamma_k \sim \text{Bern}(\pi_k)$ with $\pi_k = 0.5$, $\forall k \in \{1, ..., p\}$,

3. $c_k \sim \text{Ber}(q_k)$, with $q_k = 0.5$, $\forall k \in \{1, ..., p\}$.

(34)

4. $\sigma^2 | \delta_1, \delta_2 \sim \text{Inv-Gamm}(\delta_1, \delta_2)$, $\delta_1 = \delta_2 = 0.01$. 

23
3.3 Full conditional posterior distribution and its computation

The joint posterior distribution of the unknown parameters and hidden variables, given the data, is proportional to the product of the joint prior and the likelihood, and the fully conditional posteriors of the parameters can be easily determined by selecting the terms including the parameter in question from the joint posterior. Here the joint posterior distribution of collection of parameters is,

\[
p(\beta, \sigma^2, \gamma, c|Y, X, b) \propto p(Y|X, \gamma, \beta, \sigma^2)p(\beta|X, \gamma, \sigma^2, c, b)p(\sigma^2).
\]  

The fully conditional posterior densities of all unknowns are needed for implementing the Gibbs sampling. Each of the full conditional densities can be obtained by regarding all other parameters as known. We need to derive full condition posterior distribution for parameters.

**Full conditional posterior** \( p(\gamma|Y, X, \gamma_{(−k)}, \beta, \sigma^2, c, b) \):

The vector \( \gamma \) is obtained component wise by sampling each \( \gamma_k \) consecutively (and preferably in random order) from the Bernoulli distribution. Each component in the vector \( \gamma \) will be sampled sequentially. Because the individual \( \gamma_k \)'s are conditionally independent given \( \pi \), the update step is very similar to the univariable case. We compare the case where the \( k^{th} \) element of \( \beta \) is zero \( (\gamma_k = 0) \) against the case where it is not zero \( (\gamma_k = 1) \). The other indicator variables, call them \( \gamma_k \), are whatever their current sample is. Therefore, we need to compute the probability with which \( \gamma_k = 1 \) compared to \( \gamma_k = 0 \), given the same values for \( \gamma_k \).

We let \( \gamma_{(−k)} \) denote the vector of \( \gamma \) without \( \gamma_k \) i.e. \( \gamma_{(−k)} = (\gamma_1, \ldots, \gamma_{−1}, \gamma_{k+1}, \ldots, \gamma_p) \) and for which \( \beta_k = 0 \) if \( \gamma_k = 0 \).

The model tells that \( \gamma_k \) is independent of \( Y \) given \( \beta \) this means we can expand it as
following:

\[
Pr(\gamma_k|Y, X_γ, \gamma(-k), \beta_γ, \sigma^2, c, b) \propto p(Y|X_γ, \beta_γ, \gamma(-k), \sigma^2) p(\beta_γ|X_γ, \gamma(-k), \sigma^2, b) \\
\propto p(\beta_γ|X_γ, \gamma(-k), \sigma^2, b) p(\sigma^2)
\]  

(36)

We draw \( \gamma_k \) from a Bernoulli and we repeat this procedure for all \( k = [1, ..., p] \) predictors. Let’s tackle the cases where \( \gamma_k = 0 \) and \( \gamma_k = 1 \) in turn. Recall when \( \gamma_k = 1 \),

\[
Pr(\gamma_k = 1|Y, X_γ, \gamma(-k), \beta_γ, \sigma^2, c, b) = \frac{Pr(\gamma_k = 1|Y, X_γ, \gamma(-k), \beta_γ, \sigma^2, c, b) + Pr(\gamma_k = 0|Y, X_γ, \gamma(-k), \beta_γ, \sigma^2, c, b)}{p(\beta_γ|\gamma_k = 1, X_γ, \gamma(-k), \sigma^2, c, b)} \cdot p(\sigma^2)
\]

\[
= \frac{p(\beta_γ|\gamma_k = 1, X_γ, \gamma(-k), \sigma^2, c, b) + p(\beta_γ|\gamma_k = 0, X_γ, \gamma(-k), \sigma^2, c, b)}{p(\beta_γ|\gamma_k = 1, X_γ, \gamma(-k), \sigma^2, c, b)} \cdot p(\sigma^2)
\]

where we could cancel terms that were common to both the numerator and denominator.

Let’s tackle the cases where \( \gamma_k = 0 \) and \( \gamma_k = 1 \) in turn;

\[
l_1 = \log(p(\beta_γ|\gamma_k = 1, X_γ, \gamma(-k), \sigma^2, c, b)) = \log((2\pi)^{-p/2}|\sigma^2 G_γ X_γ^T X_γ|^{-1})^{-1/2} \exp\{-\frac{1}{2} \beta_γ^T (\sigma^2 G_γ X_γ^T X_γ)^{-1} \beta_γ\}
\]

(38)

Simplifying the first terms

\[
-\frac{1}{2} \log |G_γ X_γ^T X_γ|^{-1} G_γ = -\frac{1}{2} \log \{|G_γ|^2 |X_γ^T X_γ|^{-1}\}
\]

(39)

\[
= -\frac{1}{2} \log \left\{ \prod_{k \neq k'} g_k^2 |X_γ^T X_γ|^{-1} \right\} = -\sum_{k \neq k'} \log g_k' - \log g_k + \frac{1}{2} \log |X_γ^T X_γ|
\]

Simplifying the second term by highlighting information related to the \( k^{th} \) variable, we have

\[
\exp\{-\frac{\beta_γ^T G_γ^{-1} (X_γ^T X_γ) G_γ^{-1} \beta_γ}{2\sigma^2}\} \propto \exp\{-\frac{\beta_k g_k^{-1} (\sum_{k' \neq k} X_γ^T X_γ[k', k] \beta_k' g_k'^{-1}) + \beta_k^2 g_k^{-2} (X_γ^T X_γ)[k, k]}{2\sigma^2}\}
\]
Then the natural logarithm of the conditional posterior of $\gamma_k$ with respect to the $k^{th}$ variable when $\gamma_k = 1$ is

$$l_1 = \log \{ p(\beta_Y | \gamma_k = 1, X_Y, \gamma_{(-k)}, \sigma^2, c, b) \}$$  \hspace{1cm} (40)$$

$$\propto \log \{ |\sigma^2 G_Y (X_Y^T X_Y)^{-1} G_Y|^{-1/2} \exp \{ -\frac{1}{2} \beta_Y^T (\sigma^2 G_Y (X_Y^T X_Y)^{-1} G_Y)^{-1} \beta_Y \} \}$$

$$\propto - \sum_{k \neq k'} \log g_{k'} - \log \{g_k\} + \frac{1}{2} \log \{|X_Y^T X_Y|\} - \frac{\beta_k g_k^{-1} \sum_{k \neq k'} (X_Y^T X_Y)[k,k] g_{k'}^{-1} \} + \beta_k^2 g_k^{-2} (X_Y^T X_Y)[k,k]$$

And for $\gamma_k = 0$,

$$l_0 = \log \{ p(\beta_Y | \gamma_k = 0, X_Y, \gamma_{(-k)}, \sigma^2, c, b) \}$$  \hspace{1cm} (41)$$

$$\propto - \sum_{k \neq k'} \log g_{k'} - \log \{g_k\} + \frac{1}{2} \log \{|X_Y^T X_Y|\}$$

The natural logarithm of the conditional posterior of $\gamma_k$ is $\frac{\exp(l_1)}{\exp(l_1) + \exp(l_0)}$. We note that:

$$\gamma_k | Y, X_Y, \beta_Y, \gamma, \sigma^2, \gamma_{(-k)}, b, c \sim Bern \left( \frac{\exp(l_1)}{\exp(l_1) + \exp(l_0)} \right)$$

We use stochastic search method to find the most likely model, start with $\gamma^0 = (\gamma_1^0, \ldots, \gamma_p^0)$ and use algorithm Gibbs Sampler to calculate $\gamma^1 = (\gamma_1^1, \ldots, \gamma_p^1)$ then by iteration choose model with the highest posterior probability.

**Full conditional posterior** $p(c | X_Y, \beta_Y, \gamma, \sigma^2, c_{(-k)}, b)$

The model tells that $c_k$ is independent of $Y$ given $\beta$ then we drop any terms that does not involve $c_k$ i.e. $c_{(-k)} = (c_1, c_{k-1}, c_{k+1}, \ldots, c_p)$ and expand it as following:

$$Pr(c_k | Y, X_Y, \beta_Y, \gamma, c_{(-k)}, \sigma^2, b) \propto p(Y | X_Y, \beta_Y, c_k, c_{(-k)}, \sigma^2) p(\beta_Y | X_Y, \beta_Y, \gamma, \sigma^2, c_{(-k)}, b, c_k)$$

$$\propto p(\beta_Y | X_Y, \beta_Y, \gamma, \sigma^2, c_{(-k)}, b, c_k) p(\sigma^2)$$

(42)

We draw $c_k$ from a Bernoulli and we repeat this procedure for all $k = [1, \ldots, p]$ predictors.
Recall when $c_k = 1, g_k = g_l$ and $g_l \to \infty$ as $n \to \infty$.

\[
Pr(c_k = 1|Y, X^T, \beta, \gamma, c(-k), \sigma^2) = \frac{p(\beta|X, \beta, \gamma, \sigma^2, c(-k), b, c_k = 1)p(\sigma^2)}{p(\beta|X, \beta, \gamma, \sigma^2, c(-k), b, c_k = 1)p(\sigma^2) + p(\beta|X, \beta, \gamma, \sigma^2, c(-k), b, c_k = 0)p(\sigma^2)}
\]

Simplifying the first term when $c_k = 1$, $g_k = g_l$ and $g_l \to \infty$ as $n \to \infty$,

\[
|G_\gamma(X^T X_\gamma)^{-1}G_\gamma|^{-1/2} = \{|G_\gamma|^2|X^T X_\gamma|^{-1}\}^{-1/2} = |G_\gamma|^{-1}|X^T X_\gamma|^{1/2}
\]

\[
= \{ \left( \prod_{k\neq k'} g_k^2 \right) g_k^2 |X^T X_\gamma|^{-1} \}^{-1/2}
\]

\[
- \frac{1}{2} \log |G_\gamma(X^T X_\gamma)^{-1}G_\gamma| = - \frac{1}{2} \log \{|G_\gamma|^2|X^T X_\gamma|^{-1}\}
\]

\[
= - \frac{1}{2} \log \{ \left( \prod_{k\neq k'} g_k^2 \right) g_k^2 |X^T X_\gamma|^{-1} \} = - \sum_{k\neq k'} \log g_k - \log g_k + \frac{1}{2} \log |X^T X_\gamma|
\]

Simplifying the second term by highlighting information related to the $k^{th}$ variable, we have

\[
\exp \left\{ - \frac{\beta_{\gamma}^T G_\gamma^{-1}(X^T X_\gamma) G^{-1}_{\gamma} \beta_{\gamma}}{2\sigma^2} \right\} \propto \exp \left\{ - \frac{\beta_k g_k^{-1} \left( \sum_{k \neq k'} X^T X_{\gamma}[k', k] \beta_{k'} g_{k'}^{-1} \right) + \beta_k^2 g_k^{-2} (X^T X_\gamma)[k, k]}{2\sigma^2} \right\}
\]

Then the natural logarithm of the conditional posterior of $c_k$ with respect to the $k^{th}$ variable and by assuming $\gamma_k = 1$ then $\beta_k \neq 0$, we have

\[
l_1 = \log \{ p(\beta|X, \beta, \gamma, \sigma^2, c(-k), b, c_k = 1) \}
\]

\[
\propto - \sum_{k \neq k'} \log g_k - \log g_l + \frac{1}{2} \log \{|X^T X_\gamma|\} - \frac{\beta_k g_l^{-1} \left( \sum_{k \neq k'} X^T X_{\gamma}[k', k] \beta_{k'} g_{k'}^{-1} \right) + \beta_k^2 g_l^{-2} (X^T X_\gamma)[k, k]}{2\sigma^2}
\]

27
And for \( c_k = 0, g_k = g_s \)

\[
l_0 = \log \{ p(\beta|Y, \beta, \gamma, \sigma^2, c_{(-k)}, b, c_k = 0) \}
\]

\[
\propto - \sum_{k \neq k'} \log g_{k'} - \log g_s + \frac{1}{2} \log \{ |X^T_k X_k| \} - \beta g_{s-1} \left( \sum_{k \neq k'} X^T_k X_k[k', k] \beta_{k'} g_{k'}^{-1} \right) + \beta_{k} g_{s-2} (X^T_k X_k)[k, k] \quad \frac{2\sigma^2}{2}
\]

The conditional posterior of \( c_k \) is

\[
c_k = 1|Y, X\gamma, \beta, \gamma, c_{(-k)}, \sigma^2, b \sim Bern\left( \frac{\exp(l_1)}{\exp(l_1) + \exp(l_0)} \right) \quad (II)
\]

**Full conditional posterior** \( p(\beta|Y, X\gamma, \sigma^2, \gamma, G\gamma) :\)

We start by writing:

\[
p(\beta|Y, X\gamma, \sigma^2, \gamma = 1, G\gamma) = \frac{p(Y|X\gamma, \beta, \sigma^2, \gamma = 1, G\gamma)p(\beta|\gamma = 1, \sigma^2)p(\sigma^2)}{\int p(Y|X\gamma, \beta, \sigma^2, \gamma = 1, G\gamma)p(\beta|\gamma = 1, \sigma^2)p(\sigma^2)d\beta}\n\]

\[
= \frac{p(Y|X\gamma, \beta, \sigma^2, \gamma = 1, G\gamma)p(\beta|\gamma = 1)}{\int p(Y|X\gamma, \beta, \sigma^2, \gamma = 1, G\gamma)p(\beta|\gamma = 1)d\beta} \quad (48)
\]

where we again write the normalizing constant as \( Z \). Expanding, we get:

\[
p(\beta|Y, X\gamma, \beta, \sigma^2, \gamma = 1, G\gamma) = \frac{1}{Z} \left( 2\pi \sigma^2 \right)^{-n/2} \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta X_i)^2 \right) \left( 2\pi \sigma^2 \right)^{-1/2} \exp\left( -\frac{1}{2\sigma^2} \beta_{\gamma}^2 \right)
\]

We could write the prior over all \( \beta \)'s as a multivariate Gaussian with a diagonal covariance matrix. With a Gaussian likelihood, this prior is conjugate, such that the conditional posterior on the regression weights \( \beta \) is a multivariate Gaussian distribution.

The idea is drop any terms that does not involve \( \beta_{\gamma} \) from the joint post. dist. then write:

\[
p(\beta_{\gamma}|Y, X\gamma, \gamma, \sigma^2, G\gamma) \propto p(Y|\beta, X\gamma, \gamma, \sigma^2) p(\beta|X\gamma, \gamma, \sigma^2, G\gamma) \quad (49)
\]

\[
\propto MVN(Y|X\gamma \beta_{\gamma}, \sigma^2 I) \times MVN(\beta_{\gamma}|0, \sigma^2 G\gamma(X^T \gamma X\gamma)^{-1} G\gamma)
\]
Again writing the normalizing constant as $Z$, we expand:

\[
p(\beta|Y, \gamma = 1, c, \sigma^2) = \frac{1}{Z} |\sigma^2 I|^{-1/2} \exp\{-\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta)\} \times
\]

\[
\propto \exp\{-\frac{1}{2\sigma^2} (Y^T Y - 2X^T Y\beta + \beta^T X^T X\beta)\} \times \exp\{-\frac{1}{2\sigma^2} \beta^T (G_\gamma (X_\gamma^T X_\gamma)^{-1} G_\gamma)^{-1} \beta\}
\]

Univariate normal is special case of the multivariate normal with a one-dimensional mean

“vector” and a one-by-one variance “matrix.” Just as the probability density of a scalar normal is

\[
p(\beta) = (2\pi\sigma^2)^{-1/2} \exp\{-\frac{1}{2\sigma^2} (\beta - \mu)^2\}
\]

the probability density of the multivariate normal is

\[
p(\beta) = (2\pi)^{-n/2} (det(\Sigma_\beta))^{-1/2} \exp\{-\frac{1}{2} (\beta - \mu)^T (\Sigma_\beta)^{-1} (\beta - \mu)\}
\]

We know Normal Distribution is conjugate prior for a Normal Distribution. Thus we are looking

for a square in the form:

\[
\frac{1}{2} (\beta - \mu_\beta)^T (\Sigma_\beta)^{-1} (\beta - \mu_\beta) = \frac{1}{2} \beta^T (\Sigma_\beta)^{-1} \beta - \beta^T (\Sigma_\beta)^{-1} \mu_\beta + \frac{1}{2} \mu_\beta^T (\Sigma_\beta)^{-1} \mu_\beta
\]

Since $\mu_\beta$ does not contain a $\beta$ term then we can ignore the last term as proportional.

\[
\frac{1}{2} (\beta - \mu_\beta)^T (\Sigma_\beta)^{-1} (\beta - \mu_\beta) \propto \frac{1}{2} \beta^T (\Sigma_\beta)^{-1} \beta - \beta^T (\Sigma_\beta)^{-1} \mu_\beta
\]

\[
\propto \frac{1}{2} \beta^T [\frac{1}{\sigma^2} (X_\gamma^T X_\gamma + (G_\gamma (X_\gamma^T X_\gamma)^{-1} G_\gamma)^{-1})] \beta - \frac{1}{\sigma^2} X_\gamma^T Y\beta
\]
Let $\Sigma = \left[ \frac{1}{\sigma^2} (X^T X + G^{-1} X^T X G^{-1}) \right]^{-1}$, and $-\beta^T (\Sigma)^{-1} \mu = -\frac{1}{\sigma^2} X^T Y \beta$, then $\mu = \frac{\Sigma Y}{\sigma^2}$. 

Thus, for all $k \in \{1, \ldots, p\}$ we draw all $\beta_k$’s from:

$$
\beta_{Y_j} | X, \sigma^2, Y, \gamma, G, \gamma \sim N(\mu, \Sigma), \quad (\text{III})
$$

where $\mu = (\Sigma / \sigma^2) X^T Y$, $\Sigma = \sigma^2 (X^T X + G^{-1} X^T X G^{-1})^{-1}$ then set the $\beta_k$’s to zero for which $\gamma_k = 0$.

**Full conditional posterior** $p(\sigma^2 | Y, X, \beta, \gamma)$:

We drop any terms that do not involve $\sigma^2$ from the joint post.dist. then expand

$$
p(\sigma^2 | Y, X, \beta, \gamma) = \frac{p(Y | \beta, X, \gamma, \sigma^2) p(\beta | \sigma^2)}{\int p(Y | \beta, X, \gamma, \sigma^2) p(\beta | \sigma^2) d\sigma^2} \quad (55)
$$

To make the notation less cluttered, we will call the normalizing constant in this and all following derivations $Z$ then we have:

$$
p(\sigma^2 | Y, X, \beta, \gamma) = \frac{1}{Z} p(Y | \beta, X, \gamma, \sigma^2) p(\beta | \sigma^2) \quad (56)
$$

$$
= \frac{1}{Z} (2\pi \sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_i X_i)^2\right) \frac{\delta_1^2}{\Gamma(\delta_1)} (\sigma^2)^{-\delta_1} \exp\left(\frac{2\delta_2}{\sigma^2}\right)
$$

absorbing terms that do not depend on $\sigma^2$ into $Z$, and putting terms together — we write:

$$
p(\sigma^2 | Y, X, \beta, \gamma) = \frac{1}{Z} (\sigma^2)^{-(\delta_1 + n/2) - 1} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_i X_i)^2\right) \exp\left(-\frac{\delta_2}{\sigma^2}\right)
$$

$$
= \frac{1}{Z} (\sigma^2)^{-(\delta_1 + n/2) - 1} \exp\left(-\frac{1}{\sigma^2} \left[ \delta_2 + \frac{\sum_{i=1}^{n} (Y_i - \beta_i X_i)^2}{2} \right] \right) \quad (57)
$$

which is again an inverse Gamma distribution:

$$
\sigma^2 | Y, X, \beta, \gamma \sim Inv - Gama(\delta_1 + n/2, \delta_2 + \frac{\sum_{i=1}^{n} (Y_i - \beta_i X_i)^2}{2}) \quad (58)
$$
The conditional posterior on $\sigma^2$ with $p > 1$ predictors changes only slightly:

$$\sigma^2|Y, X, \beta, \gamma \sim \text{Inv-Gamma}(\delta_1 + n/2, \delta_2 + (Y - \beta X)^T(Y - X\beta))$$  \hspace{1cm} (IV)

### 3.4 Monte Carlo simulation

In the framework of the Bayesian approach to statistical estimation, unknown parameters of the model are regarded as random variables, with a prior distribution of probability (specified by the analyst). The Bayesian method yields the posterior distribution, i.e. the conditional probability of possible values of the parameters given the observed data. As a rule, the posterior distribution cannot be reasonably obtained by direct computation. A sample representing the Bayesian posterior distribution is obtained by applying the Markov chain Monte Carlo procedure, namely Gibbs algorithms.

The joint posterior distribution of collection of parameters is,

$$p(\beta, \sigma^2, \gamma, c|Y, X, b) \propto p(Y|X, \beta, \gamma, \sigma^2)p(\beta|X, \gamma, \sigma^2, c, b)p(\sigma^2).$$  \hspace{1cm} (59)

The Gibbs sampler is implemented by sampling repeatedly from the fully conditional posterior distributions of parameters in the model. This method generates sequence $\gamma^1, \gamma^2, \ldots$ and $c^1, c^2, \ldots$ by applying the Gibbs sampler to the complete posterior $p(\beta, \sigma^2, \gamma, c|Y, X, b)$. This produces the full sequence of parameter

$$\beta^1, \sigma^1, \gamma^1, c^1, \beta^2, \sigma^2, \gamma^2, c^2, \ldots$$

a Markov chain converging to $p(\beta, \sigma^2, \gamma, c|Y, X, b)$ in which sequence $\gamma^1, \gamma^2, \ldots$ and $c^1, c^2, \ldots$ are embedded.

Initialization: Start with an arbitrary value $(\gamma^0, \beta^0, \sigma^0, c^0)$ where $\gamma^0 = (\gamma^0_1, \ldots, \gamma^0_p)$ and $c^0 = (c^0_1, \ldots, c^0_p)$ drawn from the Bernoulli distribution.
Iteration $t$: Given $(\beta^{(t-1)}, \sigma^{(t-1)}, \gamma^{(t-1)}, c^{(t-1)})$, generate $(\beta^t, \sigma^t, \gamma^t, c^t)$.

We start with $c^0 = (c_0^0, \ldots, c_p^0)$ and use algorithm Gibbs Sampler for variable selection to calculate $c^1 = (c_1^1, \ldots, c_p^1)$ then by iteration choose best $G_Y$.

The whole Monte Carlo simulation is given in Algorithm 1.

**Algorithm 1** Gibbs algorithm for a Two-Component G prior in Variable Selection

1: procedure data matrix $X$, response $Y$, initial values, $\gamma^0, \beta^0, (\sigma^2)^0$, and $c^0$
2: for $t = 1$ to $T$ do
3: Update $\gamma$ based on $p(\gamma|Y, X, c^{t-1}, (\sigma^2)^{t-1})$ as in I
4: Update $c^t$ based on $p(c^t|Y, X, \gamma^t, \beta^t, (\sigma^2)^{t-1})$ as in II
5: Update $\beta^t$ based on $p(\beta^t|Y, X, \gamma^t, c^t, (\sigma^2)^{t-1})$ as in III
6: Update $(\sigma^2)^t$ based on $p((\sigma^2)^t|Y, X, \gamma^t, c^t, \beta^t)$ as in IV
7: end

### 3.5 Calibration of a tuning parameter

Miller\textsuperscript{2002} mentions the use of adding artificial variables to the original set of predictor variables and then checking the first appearance of them in the selected model. Wu et al.\textsuperscript{2007} develop a new approach to variable selection based on monitoring and controlling the number of artificial variables which is called pseudo-variables added to the real set of explanatory variables. They augment the set of original variables with a number of pseudo-variables that have no relationship with the responses by generation. Intuitively, a good selection criterion should not include too many of the pseudo-variables in the model selected. If a procedure selects a model containing a lot of pseudo-variables, then this procedure is too generous at including variables that should be excluded, and hence the selection procedure tends to over-fit. On the other hand, if a procedure never selects pseudo-variables, then the selection procedure is too “ruthless” at excluding variables and will seldom include weak important variables for which the regression coefficients are small but non-zero. Hence it tends to under-fit the model. As a result, the number of pseudo-variables
falsely selected is an indicator of model over-fitting or under-fitting for any selection procedure. For convenience, let us define the following notation.

\[ K_T = \text{the total number of original predictor variables.} \]
\[ K_I = \text{the number of important original predictor variables (} \beta \neq 0 \text{).} \]
\[ K_U = \text{the number of unimportant original predictor variables (} \beta = 0 \text{).} \]
\[ K_P = \text{the number of generated pseudo-variables.} \]
\[ X_1, ..., X_{K_T} = \text{original predictor variables, not including the intercept.} \]
\[ Z_1, ..., Z_{K_P} = \text{generated pseudo-variables.} \]

Obviously, \( K_I + K_U = K_T \). Here, \( K_T \) and \( K_P \) are known, and \( K_I \) and \( K_U \) are unknown. We create a set of \( K_P \) pseudo-variables that are independent of the response variable. Generation of the pseudo-variables will be discussed later.

A Bayesian false-model selection rate (BFSR) is defined as the relative frequency of selecting models that contain unimportant variables during the MCMC process. For a given data set \((Y; X)\) and selection method, the fraction of falsely-selected real predictors \( R(Y, X) \) is defined as \( R(Y, X) = \frac{K_U}{1 + K_I + K_U} \), or alternatively via the equation \( K_U - \{1 + K_I + K_U\} R(Y, X) = 0 \).

We add 1 to the sum \( K_I + K_U \) primarily because most models have intercepts, but also because it avoids special cases needed to accommodate division by zero. In general the false selection fraction is not unique and we seek the largest model consistent with controlling false selection in order to include as many important variables as possible.

Suppose that \( Z_1, ..., Z_{k_P} \) are pseudo explanatory variables, randomly generated to be independent of the response variable \( Y \), so that

\[
E(Y | X, Z) = g(\beta_1 X_1 + ... + \beta_{k_T} X_{k_T} + \beta_{k_T+1} Z_1 + ... + \beta_{k_T+k_P} Z_{k_P}) = g(\beta_1 X_1 + ... + \beta_{k_T} X_{k_T})
\]

In the regression of \( Y \) on \((X, Z)\), \( \beta_{k_T+1} = ... = \beta_{k_T+k_P} = 0 \) and thus \( Z_1, ..., Z_{k_P} \) are uninformative variables. Now consider applying a variable selection method repeatedly to \( L \) randomly generated, augmented data sets, \((Y; X, Z_l)\), where \( Z_l \) are \( n \times k_P \) replicate pseudo-predictor variables.
matrices, \( l = 1, \ldots, L \). Over fitting is indicated if the method selects a high percentage of the \( k_P \) phony predictor variables on average; underfitting is indicated if the method selects very few or no phony variables on average. Once the underfitting or overfitting tendency of a selection method is revealed, the method can often be tuned to reduce undesirable fitting tendencies. The objective is to generate pseudo variables so that the average inclusion probabilities of informative variables are approximately equal with data \((Y; X)\) and \((Y; X, Z)\); and the average inclusion probabilities of uninformative variables (real and pseudo) are approximately equal with data \((Y; X)\) and \((Y; X, Z)\). Here, the \( n \) rows of entries in the \( n \times k_P \) matrix \( Z \) are obtained by randomly permuting the rows of \( X \) which restricts \( k_P = k_T \). The variants are such that the pseudo predictors have sample means and sample correlations with the real predictors that are identically equal to zero. Thus the variants work like Monte Carlo swindles to reduce variation.

Now we introduce a tuning parameter to optimize the prior. This tuning parameter is adaptively selected with the help of carefully created pseudo-variables. Compared to the existing methods, a major contribution of this study exists in the extension of the classical g-prior and the inclusion of pseudo-variables to optimize the extended g-prior. If we denote the proportion of pseudo-variables among the selected variables by \( BFSR = \frac{T_U}{T} \), where \( T \) is the number of MCMC iterations, and \( T_U \) models with unimportant variables, it will quantify the extent of falsely selected pseudo-variables. To reduce the variation in \( BFSR \), we repeatedly generate sets of pseudo-variables \( T \) times. The estimation of BFSR satisfies the following two properties:

(A1) states that real unimportant variables and phony unimportant variables have the same probability of being selected on average; that is, \( E\{I_P(b)\} = E\{I_U(b)\} \) then

\[
E(BFSR) = E(T_P/T),
\]

where \( T_P \) is the number of models with pseudo variables included among the \( T \) iterations i.e. \( BFSR = T_P/T \)

(A2) states that the real important variables have the same probability of being selected whether phony variables are present or not.
In this vein we see that (A1) and (A2) indicate that the phony variables should behave as much like the real unimportant variables as possible (A1), and should have as little impact as possible on the selection rate of the real important variables (A2).

For start point we permute the rows of $X$ design matrix and denote new matrix by $P_X$. Permutation produces pseudo variables that when appended to the real data create what are essentially matched pairs. To each real variable there corresponds a pseudo variable with identical sample moments and also with preservation of correlations. In fact in the all-null model (no real important variables), the permutation method ensures that (A1) holds asymptotically. The pseudo-variables are then formed by the residuals through regressing $P_X$ on $X$. Specifically, the values of the pseudo-variables are given by $(I - H_X)P_X$ where $H = X(X^TX)^{-1}X^T$. Projection via $(I - H_X)$ helps reduce the effect of the pseudo variables on the selection probabilities of the real important variables, thus lessening the extent to which (A2) is violated. After introducing the pseudo-variables, data $(Y; X)$ is augmented to $(Y; X, R_P)$ with $R_P = (I - H_X)P_X$.

The proposed adaptive selection process is motivated by the work of Browne and Draper\textsuperscript{2006}, in which an adaptive approach is introduced at the beginning of MCMC posterior sampling to select variances for jumping proposals. In this work, we use this idea to select $b$.

From starting values based on the estimated $b$, we first employ a Bayesian false-model selection rate (but with an upper bound $w$, the target and the initial value of $b$ is $b_0$) during $b$ are adaptively tuned and eventually fixed for the remainder of run; this is followed by the usual burn-in period; and then the main monitoring run from which posterior summaries are calculated occurs. Usually $b_0$ takes a relatively small value to potentially include more variables. The tuning of $b$ is based on achieving a Bayesian false-model selection rate $w$ for each parameter that lies within a specified tolerance interval $(w - \delta, w + \delta)$.

The algorithm examines empirical acceptance rates in batches of $T$ iterations, comparing them for each parameter with the tolerance interval and modifying the $BF^SR$ before going on to the next batch of $T$. This modifies the proposal standard deviation by a greater amount the
farther the empirical acceptance rate is from the target \( w \). If \( \hat{BF}_{SR} \) is too low, the proposed moves are too big, so \( b_m \) is decreased; if \( \hat{BF}_{SR} \) is too high, the parameter space is being explored with moves that are too small, and \( b_m \) is increased. If the \( \hat{BF}_{SR} \) values are within the tolerance interval during three successive batches of \( T \) iterations, the parameter is marked as satisfying its tolerance condition, and once all parameters have been marked the overall tolerance condition is satisfied and adapting stops. After a parameter has been marked it is still modified as before until all parameters are marked, but each parameter only needs to be marked once for the algorithm to end. After a parameter has been marked it is still modified as before until all parameters are marked. The detailed procedure for finding the tuning parameter is summarized as follows in Algorithm 2.

**Algorithm 2 MCMC Simulation Process for the Adaptive Calibration of \( b \)**

1. procedure data matrix \( X \), response \( Y \), \( b_0 \)
2. for \( m = 0 \) to \( M \) do
3. Fix \( b_m \)
4. Calculate \( P_X \) by permuting the rows of design matrix \( X \)
5. Generate \( X_{pseudo} \) by the residuals through regression \( P_X \) on \( X \). i.e. \( X_{pseudo} = (I - X(X^TX)^{-1}X^T)P_X \)
6. Fit regression model using \( Y \) and \( X_{all} = (X, X_{pseudo}) \)
7. for \( t = 1 \) to \( T \) do
8. Drawn posterior samples and generate \( \beta_t^i, (\sigma^2)^t, \gamma_t, c_t \), based on in Algorithm 1
9. Calculate \( \hat{BF}_{SR} = \sum_{t=1}^{T} I_t \), where \( I_t = 1 \), if at least one pseudo variable is selected, and \( I_t = 0 \) otherwise for \( t = 1, \ldots, T \)
10. End for
11. if \( \hat{BF}_{SR} => w + w_0 \) then
12. \( b_{m+1} \rightarrow b_m [2 - \frac{1-\hat{BF}_{SR}}{1-w}] \)
13. else \( b = b_m \).
14. End for

If \( BF_{SR} \) is greater than 0.5 then the number of the unimportant variables that we have found is more than real ones therefore we increase the distance between variables. If \( BF_{SR} \) is less than 0.5 then the number of unimportant variables are less than real number ones that we have to decrease the distance. After obtaining \( b \), one can then fit Gibbs simulation by excluding pseudo-variables.
Briefly we proposed a method based on the simple idea that the tendency of a variable selection method to overfit or underfit can be revealed by the use of pseudo explanatory variables. The intuition behind this algorithm is that the pseudo-variables can be used to assess the tuning parameter selected. Therefore, one can choose a tuning parameter that controls the percentage of pseudo-variables in the selected model.

3.6 Properties of the $G$-prior

Bayes factors (BFs) are indices of relative evidence of one “model” over another, which can be used in the Bayesian framework as alternatives to classical (frequentist) hypothesis testing indices (such as p-values). According to Bayes’ theorem:

$$P(M|D) = \frac{P(D|M) \times P(M)}{P(D)}$$  \hspace{1cm} (61)

Then by comparing two models, we get:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)} \times \frac{P(M_1)}{P(M_2)}$$  \hspace{1cm} (62)

Where the middle term is the Bayes factor:

$$BF_{12} = \frac{P(D|M_1)}{P(D|M_2)}$$  \hspace{1cm} (63)

Thus, Bayes factors can be seen either as a ratio quantifying the relative likelihood of two models in light of some observed data as they can be computed by comparing marginal likelihoods, or as the degree by which some prior beliefs about the relative odds of two models are to be updated as they can be computed by dividing posterior odds by prior odds, as we will soon demonstrate. A ‘large’ value of $BF_{12}$ indicates support for $M_1$ relative to $M_2$, and a ‘small’ value (> 0) indicates support for $M_2$ relative to $M_1$. Bayes factor consistency
refers to the stochastic convergence of $BF_{12}$, under the true probability distribution, such that $BF_{12} \to \infty$ if $M_1$ is the best model, and $BF_{12} \to 0$ if $M_2$ is the best model.

The following Proposition states that, compared to the ratio of posterior model probability formulated in Zellner’s g-prior, the one built upon two-component $G$ converges in probability to zero more quickly for any model $M_u$ other than the true model $M_t$. But before to prove the Proposition, we need the following lemma:

**Lemma**. Assume $X^T X / n \to \Sigma_X$ as $n \to \infty$ with $\Sigma_X$ being the covariance of $X$ and $\Sigma_X$ positive definite. As $n \to \infty$, given all other parameters, $c_k = 0$ with probability 1 if $\beta_k \neq 0$.

**Proof**: We want to show that $\lim_{n \to \infty} Pr(c_k = 0|X, \beta, \gamma, \sigma^2, c(-k), b) = 1$.

Recall the natural logarithm of the conditional posterior of $c_k$:

$$Pr(c_k|Y, X, \beta, \gamma, c(-k), \sigma^2, b) \propto p(Y|X, \beta, \gamma, c(-k), \sigma^2)p(\beta|X, \beta, \gamma, \sigma^2, c(-k), b, c_k)$$

$$\propto p(\beta|X, \beta, \gamma, \sigma^2, c(-k), b, c_k)p(\sigma^2)$$  \(64\)

Let’s tackle the cases where $c_k = 0$ and $c_k = 1$ in turn. Recall when $c_k = 1$,

$$p(\beta|X, \beta, \gamma, \sigma^2, c(-k), b, c_k = 1) \propto |\sigma^2 G(Y X)^{-1} G_Y|^{-1/2} \exp\left\{ -\frac{1}{2} \beta^T (\sigma^2 G(Y X)^{-1} G_Y)^{-1} \beta \right\}$$

Simplifying the first terms when $c_k = 1, g_k = g_l$ and $g_l \to \infty$ as $n \to \infty$,

$$|G(Y X)^{-1} G_Y|^{-1/2} = \{|G_Y^2|X Y^{-1}|X Y|^{-1}\}^{-1/2} = |G_Y|^{-1}|X Y|^{1/2}$$ \(66\)

$$= \left\{ \left( \prod_{k \neq k'} g_k^2 |X Y|^{-1} \right) \right\}^{1/2}$$

$$-\frac{1}{2} \log |G(Y X)^{-1} G_Y| = -\frac{1}{2} \log \{|G_Y^2|X Y^{-1}|X Y|^{-1}\}$$ \(67\)

$$= -\frac{1}{2} \log \left\{ \left( \prod_{k \neq k'} g_k^2 |X Y|^{-1} \right) \right\} = -\sum_{k \neq k'} \log g_k - \log g_k + \frac{1}{2} \log |X Y|$$
Simplifying the second term by highlighting information related to the $k^{th}$ variable, we have
\[
\exp\left\{-\frac{\beta_Y g^{-1}}{2\sigma^2} (X_T^TY)G^{-1}_Y \beta_Y \right\} \approx \exp\left\{-\frac{\beta_k g^{-1}_k (\sum_{k\neq k'} X_T^TY[k',k] \beta_{k'} g^{-1}_{k'}) + \beta_k^2 g^{-2}_{k} (X_T^TY[k,k])}{2\sigma^2} \right\}
\]

based on the natural logarithm of the conditional posterior of $c_k$, assuming $\beta \neq 0$ we have
\[
\log(Pr(c_k = 1|Y, X_Y, \beta_Y, \gamma, c_{(-k)}, \sigma^2, b)) - \log(Pr(c_k = 0|Y, X_Y, \beta_Y, \gamma, c_{(-k)}, \sigma^2, b))
\]
\[
= \log(g_s/g_l) - \frac{(g_l^{-1} - g_s^{-1}) \beta_k (\sum_{k\neq k'} X_T^TY[k',k] \beta_{k'} g^{-1}_{k'})}{2\sigma^2} - \frac{(g_l^{-2} - g_s^{-2}) \beta_k^2 (X_T^TY[k,k])}{2\sigma^2}
\]

In the speed of $\log(n^{a_0})$ with $a_0 < 1$, $\lim_{n \to \infty} \log(g_s/g_l) = -\infty$. The numerator of the second term is
\[
(g_l^{-1} - g_s^{-1})n\beta_k (\sum_{k\neq k'} X_T^TY[k',k]/n\beta_{k'} g^{-1}_{k'}) = \beta_k (\sum_{k\neq k'} X_T^TY[k',k]/n\beta_{k'} (n/(g_{k'} g_l) - n/(g_{k'} g_s))
\]

Similarly, the third term is $(n/g_l^{-2} - n/g_s^{-2}) \beta_k^2 (X_T^TY[k,k])/n$. Therefore,
\[
\lim_{n \to \infty} \log(Pr(c_k = 1|Y, X_Y, \beta_Y, \gamma, c_{(-k)}, \sigma^2, b)) - \log(Pr(c_k = 0|Y, X_Y, \beta_Y, \gamma, c_{(-k)}, \sigma^2, b)) = -\infty
\]

since $g_l \to \infty$ in the order $O(n)$, $g_s \to \infty$ in the order of $O(n^\psi)$ with $1/2 < \psi < 1$, and

$(X_T^TY)/n \to \Sigma Y$ with $\Sigma Y$ positive definite, the last two terms in above equation are finite.

This implies that if variable $k$ is important, then $g_k$ takes the value of $g_l$ with probability approaching to 0 as $n \to \infty$. Consequently, if $\beta_k \neq 0$, and $c_k = 0$ that is, $g_k = g_s$, then we have
\[
\lim_{n \to \infty} Pr(c_k = 0|Y, X_Y, \beta_Y, \gamma, \sigma^2, c_{(-k)}, b) = 1
\]
which implies that if a variable $k$ takes $g_k = g_s$, then this variable is to be treated as an important variable and will be included in the model.

Following the above conclusion, It is straight forward that
\[
\lim_{n \to \infty} Pr(Y_k = 0|Y, X_Y, \beta_Y, \gamma_{(-k)}, \sigma^2, c_k = 1, c_{(-k)}, b) = 1,
\]
which implies that if a variable \( k \) takes \( g_k = g_l \), then this variable is more likely to be treated as an unimportant variable and will be excluded from the model. The choice between \( g_l \) and \( g_s \) will be determined by the data.

**Proposition**. Denote by \( M_t \) the underlying true model and \( M_u \) a model such that \( M_u \neq M_t \). Define \( R_{u,t}(g_s) = \text{Pr}(M_u|Y, X, g_s)/\text{Pr}(M_t|Y, X, g_s) \), the ratio of posterior model probabilities (\( M_u \) over \( M_t \)) under the Zellner’s g-prior with \( g_k = g_s \) for all \( k, k = 1, \ldots, p \), and \( R_{u,t}(G_γ) = \text{Pr}(M_u|Y, X, G_γ)/\text{Pr}(M_t|Y, X, G_γ) \), the ratio of posterior model probabilities using the proposed prior as opposed to the Zellner’s g-prior.

Assume \( X^T X / n \rightarrow X \) as \( n \rightarrow \infty \) with \( X \) being the covariance of \( X \) and \( X \) positive definite. As \( n \rightarrow \infty \), the ratios \( R_{u,t}(g_s) \) and \( R_{u,t}(G_γ) \) (conditional on \( X \)) satisfy \( R_{u,t}(g_s) \rightarrow 0 \), \( R_{u,t}(G_γ) \rightarrow 0 \), and \( R_{u,t}(G_γ) = o(R_{u,t}(g_s)) \).

**Proof**: The joint distribution of \( \theta_0 = (β_γ, σ^2) \), a collection of parameters for a given model determined by \( γ \), and \( Y \) conditional on \( X_γ \) and \( G_γ \) is

\[
p(θ_0, Y|X_γ, γ, G_γ) = p(β_γ|Y, X_γ, γ, σ^2, G_γ)p(σ^2|Y, X_γ, γ, G_γ)p(Y|X_γ, γ, G_γ).
\] (72)

By taking \( g_k = g_s \) in \( G_γ \), it is reduced to the setting under the Zellner’s g-prior.

The first term is the full conditional posterior distribution of \( β_γ \);

\[
(β_γ|Y, X_γ, γ, σ^2, G_γ, c, π, q) ∼ N(μ_β, Σ_β)
\] (73)

where \( μ_β = (Σ_β/σ^2)X_γ^T Y \), \( Σ_β = σ^2(X_γ^T X_γ + G_γ^{-1}X_γ^T X_γ G_γ^{-1})^{-1} \).

The second part is obtained by integrating out \( β \) and is marginal posterior distribution of \( σ^2 \) under a specific model

\[
p(σ^2|Y, X_γ, γ, G_γ) ∝ (σ^2)^{-1/2} \exp \left\{ -\frac{Y^T Q_γ Y}{2σ^2} \right\} = \text{Inv - Gam} \left( \frac{n}{2}, \frac{Y^T Q_γ Y}{2σ^2} \right),
\] (74)

where \( Q_γ = I - X_γ(X_γ^T X_γ + G_γ^{-1}X_γ^T X_γ G_γ^{-1})^{-1}X_γ^T \).
The last term is related to the calculation of Bayes factor:

\[ p(Y|X_\gamma, \gamma, G_\gamma) \propto (Y^T Q_\gamma Y)^{-n/2} \frac{|G_\gamma^{-1}(X_\gamma^T X_\gamma)G_\gamma^{-1}|^{1/2}}{|X_\gamma^T X_\gamma + G^{-1}(X_\gamma^T X_\gamma)G^{-1}|^{1/2}}. \]  \hspace{1cm} (75)

If variable \( k \) is important from previous Lemma for a given \( \sigma^2 \) under a model defined by \( \gamma \), \( c_k \) takes the value of 0 with probability 1 as \( n \to \infty \), equivalently, \( g_k = g_s \) with probability 1 if \( \beta_k \neq 0 \). Thus if \( \gamma \) specifies the true model \( t \), then \( g_k \to g_s \) in probability 1 for all \( k \) in model \( t \) as \( n \to \infty \). Thus we have \( \lim_{n \to \infty} \log(Pr(M_t|Y, X, G_\gamma)) - \log(Pr(M_l|Y, X, g_s)) = -\infty \), so

\[ \lim_{n \to \infty} (Pr(M_t|Y, X, G_\gamma)/Pr(M_l|Y, X, g_s)) = 0 \quad (A) \]

We define \( R_u(G_\gamma, g_s) \) and by Minkowski determinant theorem for given two positive definite \( n \times n \) symmetric matrices \( A, B \), \( det(A + B)^{1/n} \geq det(A)^{1/n} + det(B)^{1/n} \) we have

\[ R_u(G_\gamma, g_s) = \frac{Pr(M_u|Y, X, G_\gamma)}{Pr(M_u|Y, X, g_s)} \propto \frac{Pr(Y|X_\gamma, \gamma, G_\gamma)}{Pr(Y|X_\gamma, \gamma, g_s)} \]

\[ = \left[ \frac{Y^T Q_{\gamma, g_s} Y}{Y^T Q_{\gamma, G_\gamma} Y} \right]^{n/2} \frac{|X_\gamma^T X_\gamma|^{1/2}(1 + g_s^{-2})^{k_0/2}}{|X_\gamma^T X_\gamma + G^{-1}(X_\gamma^T X_\gamma)G^{-1}|^{1/2}} \frac{|G_\gamma^{-1}(X_\gamma^T X_\gamma)G_\gamma^{-1}|^{1/2}}{|g_s^{-1}(g_s^{-1})^{k_0-m} + G^{-1}(X_\gamma^T X_\gamma)G^{-1}|^{1/2}} \]

where \( Q_{\gamma, g_s} = Q_\gamma \) by taking all \( g_k = g_s \), \( Q_{\gamma, G_\gamma} = Q_\gamma \) by using the two-component \( G \), and we assume model \( M_u \) has \( k_0 \) variables.

Suppose,

\[ A_{g_s} = \frac{1}{1 + g_s^{-2}} X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T, \text{ and } A_{G_\gamma} = X_\gamma (X_\gamma^T X_\gamma + G_\gamma^{-1}(X_\gamma^T X_\gamma)G_\gamma^{-1})^{-1} X_\gamma^T. \]

The sum of squared errors under the model with \( g_k = g_s \) and the model with \( G_\gamma \) with the conditional posterior mean of \( \beta_\gamma \) are \( Y^T (I - A_{g_s}, A_{g_s}) Y \) and \( Y^T (I - A_{G_\gamma}, A_{G_\gamma}) Y \), respectively. If we denote the diagonal elements of \( G_\gamma / g_s \) with \( \text{diag}(1, g_s / g_l) \) then
\[ A_{G_Y} = X_Y(X_Y^T X_Y + g_s^{-2} \text{diag}(1, g_s/g_l)X_Y^T X_Y \text{diag}(1, g_s/g_l))^{-1} X_Y^T. \]

As \( n \) becomes large, \( Y^T (I - A_{g_s} A_{g_s}) Y < Y^T (I - A_{G_Y} A_{G_Y}) Y \) since \( \lim_{n \to \infty} g_s/g_l = 0 \), and then we have \( Y^T (I - A_{g_s}) Y < Y^T (I - A_{G_Y}) Y \), because \( X_Y^T X_Y \) in \( A_{g_s} \), and \( (X_Y^T X_Y + G_Y^{-1} X_Y G_Y^{-1}) \) in \( A_{G_Y} \) are positive definite. Therefore, \( \lim_{n \to \infty} \frac{Y^T Q_{gs} Y}{Y^T Q_{G_Y} Y} n/2 = 0 \), and for given \( k_0 < \infty \)

\[
\lim_{n \to \infty} R_u(G_Y, g_s) = 0 \quad (B)
\]

Finally by combining Results (A) and (B) we have

\[
\lim_{n \to \infty} \frac{R_{u,t}(G_Y)}{R_{u,t}(g_s)} = \lim_{n \to \infty} \frac{\frac{\Pr(M_u | Y, X, G_Y)}{\Pr(M_t | Y, X, G_Y)}}{\Pr(M_u | Y, X, g_s)} = 0.
\]

(78)

It is straightforward that \( \lim_{n \to \infty} R_{u,t}(G_Y) = 0 \) asymptotically since \( R_{u,t}(g_s) \) and \( R_{u,t}(G_Y) \), are Bayes factors when taking uniform priors for both models \( M_u, M_t \), and \( R_{u,t}(G_Y) = o(R_{u,t}(g_s)) \).

In summary, Proposition states that, compared to the ratio of posterior model probability formulated in Zellner’s g-prior, the one built upon two-component G converges in probability to zero more quickly for any model \( M_u \) other than the true model \( M_t \). In other words, that means the probability of selecting true model grows with sample size in both methods and the growth of \( R_{u,t}(G_Y) \) with sample size is faster than the growth of \( R_{u,t}(g_s) \).
Chapter 4

Simulation

To assess the Two-component $G$-prior, we conduct two types of studies. First through simulations, we examine the parameters of the approach. Second, we apply the proposed methods to the above simulated data sets. For each data set, two Markov chains are simulated for the purpose of convergence assessment. Through simulations, in this chapter we demonstrate the proposed variable selection approach and take $g_l = b f_1(n) = bn$, and $g_s = b f_2(n) = bn^\psi$ with $\psi = 0.55$.

4.1 Simulation Design

Example 1. The purpose of this example is to demonstrate the methods. A data set of size $n$ is generated. We generate five uncorrelated variables $X_1, ..., X_5$ from normal $N(0, I)$ and a two-level categorical variable $Z$ from binomial distribution. First variable is important with coefficients $\beta_1 = 1$ and it also interacts with a two-level categorical variable with coefficient $2$, which gives $\beta = (1, 0_{1 \times 5}, 2, 0_{1 \times 4})$. We can calculate $Y$ by the equation $Y = 1 * X_1 + 2 * X_1 * Z + error$, where error is also generated by normal distribution. The variance of the random error is $\sigma^2 = 1$, we also suppose that the initial $\beta$ comes from the equation $lm(Y \sim X - 1)$. The model matrix creates a design matrix from the description given
in terms(object). The pseudo-variables are generated based on the criteria described in Section 3.5 and the maximum number of covariates after including pseudo variables becomes 22.

**Example 2.** A data set of size \( n \) with 20 candidate variables is simulated, among the 20 variables, the first 4 are important with regression coefficients \((\beta_1, ..., \beta_4) = (1, 1.5, 2, 1.8)\). The correlation between every two candidate variables is \((0.6)^{|i-j|}, i, j = 1, ..., 20\). Other settings are the same as in Example 1. This example is to demonstrate the performance of the method when the number of covariates is relatively large and those covariates are correlated.

We consider various sample sizes \( n = 100, 200, 300, 400 \) for each example, and for each sample size, 1000 data sets were simulated. We run the Gibbs sampler for different values of \( b = 1, 50, 100, 150 \), and different number of variables; \( p = 11, 31 \).

### 4.2 Results and Discussion

We applied the proposed methods to the above simulated data sets. For each data set, two Markov chains are simulated for the purpose of convergence assessment. Each chain starts from a certain number of iterations with an initial value of \( b_0 \).

For example 1, the dimension of inputs and outputs are \( X_{100\times11}, Y_{100\times1}, \gamma_{1\times11}, \beta_{1\times11}, c_{100\times11} \) and \( \sigma^2 \). In this simulation we validate the importance of variables one (first important variable of five uncorrelated variables \( X_1, ..., X_5 \) from normal \( N(0, I) \)) and two-level categorical variable (since the interaction of the first variable with a two-level categorical variable is important) which we have already known from \( \beta = (1, 0_{1\times5}, 2, 0_{1\times4}) \). Moreover, we investigate the role of \( b \) in distinguishing between important and unimportant variables.

At the first step by a sample size of 100 for \( X \) and associated \( Y \), we run the algorithm to find the outputs: \( \gamma, \beta, \sigma^2 \) and \( c \).

We repeat this algorithm 1000 times to make the results more robust for different
samples. We conclude the estimation probability of posterior $\gamma$ by averaging over the all outputs. The result is shown for different distance in table 4.1.

According to figure 4.1 variables one and seven have the highest probability for different distances that is in complete agreement with our first assumption.

At the second step we evaluate the role of distance in differentiating between the significance of variables. The distance between the two components was adjusted by a tuning parameter $b$. The tuning parameter was selected adaptively by controlling a Bayesian false-model selection rate. We have 11 important variables and want to assess the impact of the distance between important and unimportant variables on the selection of variables. Then we generate 11 pseudo variables and add them to the model.

As figure 4.2 demonstrates, there is an inverse relation between the various distances and the estimated posterior probability selecting unimportant variables. According to scree plot in figure 4.2, by increasing the distance from 0.5 to 100, the estimated posterior probability drops from approximately 0.5 to 0.25.

The result for example 2 is shown in figure 4.3. Based on this figure the first four variables are the most important ones that matches with the assumed model. In addition, this figure indicates the reverse dependency between distance and the estimated posterior probability selecting unimportant variables same as example one.
Table 4.1: Estimated posterior probability selecting for 11 variables with sample size n=100 and various distance b=0.5, 1, 50, 100

<table>
<thead>
<tr>
<th>Principle Component</th>
<th>b=0.5</th>
<th>b=1</th>
<th>b=50</th>
<th>b=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>gamma1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gamma7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gamma4</td>
<td>0.557</td>
<td>0.501</td>
<td>0.375</td>
<td>0.323</td>
</tr>
<tr>
<td>gamma5</td>
<td>0.533</td>
<td>0.501</td>
<td>0.45</td>
<td>0.39</td>
</tr>
<tr>
<td>gamma8</td>
<td>0.525</td>
<td>0.527</td>
<td>0.403</td>
<td>0.383</td>
</tr>
<tr>
<td>gamma3</td>
<td>0.513</td>
<td>0.467</td>
<td>0.411</td>
<td>0.371</td>
</tr>
<tr>
<td>gamma9</td>
<td>0.523</td>
<td>0.493</td>
<td>0.45</td>
<td>0.401</td>
</tr>
<tr>
<td>gamma2</td>
<td>0.491</td>
<td>0.506</td>
<td>0.377</td>
<td>0.397</td>
</tr>
<tr>
<td>gamma10</td>
<td>0.495</td>
<td>0.471</td>
<td>0.445</td>
<td>0.385</td>
</tr>
<tr>
<td>gamma11</td>
<td>0.451</td>
<td>0.433</td>
<td>0.502</td>
<td>0.454</td>
</tr>
<tr>
<td>gamma6</td>
<td>0.487</td>
<td>0.461</td>
<td>0.351</td>
<td>0.325</td>
</tr>
</tbody>
</table>

Figure 4.1: Patterns of estimated posterior probability selecting on 11 principle components across different distance of b with sample size n=100
Figure 4.2: Patterns of estimated posterior probability selecting on 31 principle components across different distance of b with sample size n=200
Chapter 5

Conclusion

Variable selection has become an important challenge in statistical analysis. Statisticians are often faced with a large number of potential regressors; however, usually only a small subset of these really have an influence on the response variable. The goal of variable selection is to distinguish between zero effects and non zero effects. In the Bayesian approach a prior is assigned to regression coefficients, which is a mixture of a 'spike' distribution and a flat 'slab' distribution. The spike allows shrinkage of small effects to zero. The probability that a regressor should be included in the final model is given by the posterior inclusion probability of the regressor estimated by the proportion that the coefficient was assigned to the slab component of the prior during MCMC.

In the literature different prior proposals for spike and slab priors have been discussed, e.g. independence prior, Zellner's g-prior, fractional prior, NMIG-prior and SSVS-prior. The goal of this master thesis which is a continuation of a project by Zhang et al.\textsuperscript{2016} published in 2016, was to examine the influence of extension of the Zellner's g-prior on variable selection. Instead of using a universal single $g$ as in the classic Zellner's g-prior, we recommended the use of two different $g_k$'s (the two-component Gprior) to distinguish important variables from unimportant ones.
In this thesis we assessed this new methodology for Bayesian variable selection in multiple linear regression in the following steps.

- Extension of the original scalar $g$ to a diagonal matrix $G$ that controls the stability of the prior on the coefficients $\beta$.

- Investigation the proof of proposition in details that indicates using a two-component $G$ can be more efficient in identifying truly important variables than when a scalar $g$ is used.

- Discussion and derivation the prior and fully conditional posterior distributions of each parameter from the joint posterior distribution of parameters.

- Implementation fully conditional posterior distributions of each parameter into Gibbs sampling algorithm to complete posterior distribution.

- Assessing the impact of tuning parameter $b$, the distance between important and unimportant variables, on the selection of variables by controlling Bayesian false model selection rate with respect to unimportant variables based on creating pseudo variables.

- Summarizing and discussing our findings by simulation which validates and evaluates the proposed method.

In the simulations we generated the data sets with 100, 200 observations and 11, 31 regressors with strong/zero effects. We used 0.5, 1, 50, and 100 as distances between important and unimportant variables. The results verifies the proposed model.

Left as future work is the further extension of the two-component method. We proposed a two-component $G$ method composed of $g_s$ and $g_l$. This is based on the belief that under the framework of variable selection there are two types of variables, important and unimportant variables. This idea can be generalized for more than two categories and the Metropolis-Hastings algorithm is suggested in simulation for further study.
Appendix

R codes

The basic outline of the code for the Gibbs Sampler is Gibbsfunctions(). The section of the code devoted to updating each parameter. The data for the simulation study was generated in R using the function Datagenfunction(). This function generates the rmvnorm(n, mean.X, Sigma.X) data as discussed using the appropriate distribution functions in R. The discrete variables are generated using rbinom(n, 1, 0.5) function.

A: Gibbs functions();

\[
gibbs <- \text{function}(Y, X, b0, a1 = .01, a2 = .01, niter, nr - burnin = \text{round}(niter/4, 2))
\]

B: Data generated function();

\[
Out <- \text{mydata}(p = 5, n = 30, rho = 0), then Y, X
\]

C: Simulation study;

\[
mc.vals <- \text{gibbs}(Y, X, b0 = 50, a1 = .01, a2 = .01, niter = 1000, nr - burnin = \text{round}(1000/4, 2))
\]
# update gamma

gamma_fnc <- function(gamma_prev, X, Y, G, beta_prev){
  # res[1, seq(p)] <- c(1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1)
  # initial value for gamma
  gamma_prev <- res[1, seq(p)]

  for (j in 1:length(gamma_prev)) {
    n <- nrow(X)
    # cat("gamma_prev", gamma_prev, ",\n"
    # if (sum(as.integer(gamma_prev)) >= 2){
      Xj=X[,j]
      X_j=X[,,-j]
      X_j_gammaprod=X_j[,which(gamma_prev[-j]==1)]
      A <- t(Xj)%*%X_j_gammaprod
      # update G_j based on gamma=1
      G_j <- G[,-j]
      G_gammaprod=G[-j,which(gamma_prev[-j]==1),which(gamma_prev[-j]==1)]
      # update beta_j based on gamma=1
      beta_j <- beta_prev[-j]
      beta_gammaprod = beta_j[which(gamma_prev[-j]==1)]
      B <- beta_gammaprod %*% solve(G_gammaprod)
      D <- A%*%t(B)

      gamma_prev[j] <- 1
      gamma_prev <- gamma_prev
      X_gammaprod = X[, which(gamma_prev==1)]
      X_gammaprod <- as.matrix(X_gammaprod)
      beta_gammaprod = beta_prev[which(gamma_prev==1)]
      beta_gammaprod <- as.matrix(beta_gammaprod)
      G_gammaprod = G[which(gamma_prev==1), which(gamma_prev==1)]
      G_gammaprod <- as.matrix(G_gammaprod)
      err1_Y = X_gammaprod %*% beta_gammaprod

      sigma2_prev1 <- 1 / rgamma(1, 0.01 + n/2, 0.01 + t(err1_Y) %*% err1 / 2)
      # compute chance parameter of the conditional posterior of gamma_j (Bernoulli)
      l1 <- -(n/2)*log(2*pi*sigma2_prev1) -
      log(det(G_gammaprod)+(L/2)*log(det(t(X_gammaprod)%*%X_gammaprod))) -
      (L/2*sigma2_prev1)*((beta_prev[j]%*%solve(G)[j,j])*D +
      beta_prev[j]/2%*%solve(G)[j,j]*t(x)%*%x)[j,j] + log(0.5)
      gamma_prev[j] <- 0
      gamma_prev <- gamma_prev
      X_gammaprod = X[, which(gamma_prev==1)]
      X_gammaprod <- as.matrix(X_gammaprod)
      beta_gammaprod = beta_prev[which(gamma_prev==1)]
      beta_gammaprod <- as.matrix(beta_gammaprod)
      G_gammaprod = G[which(gamma_prev==1), which(gamma_prev==1)]
      G_gammaprod <- as.matrix(G_gammaprod)
      err0_Y = X_gammaprod %*% beta_gammaprod

      sigma2_prev0 <- 1 / rgamma(1, 0.01 + n/2, 0.01 + t(err0_Y) %*% err0 / 2)
      l0 <- -(n/2)*log(2*pi*sigma2_prev0) -
      log(det(G_gammaprod)+(L/2)*log(det(t(X_gammaprod)%*%X_gammaprod))) + log(0.5)

      # sample gamma_j from a Bernoulli and update
      gamma_prev [j] <- rbinom(1, 1, exp(l1)/(exp(l1) + exp(l0)))
  }
}
if (sum(as.integer(gamma_prev)) >= 2){
gamma_prev <- gamma_prev
}
return(gamma_prev)
}
# update X based on gamma=1
xgam_fnc <- function(gamma_prev, X){
  X_gammarange = X[, which(gamma_prev == 1)]
  return(X_gammarange)
}
# update G based on gamma=1
gamma_fnc <- function(X, gamma_prev, c_prev, b0){
g1 <- nrow(X)
gs <- sqrt(nrow(X))
g_prev <- numeric()
for (i in 1:length(c_prev)) {g_prev[i] <- c_prev[i] * b0 * g1 + (1-c_prev[i]) * b0 * gs}
G <- diag(g_prev)
gamma_gammaj = G[which(gamma_prev == 1), which(gamma_prev == 1)]
return(G_gammaj)
}
# update c
c_fnc <- function(gamma_prev, X, g, beta_prev, sigma2_prev, c_prev, b0){
n <- nrow(X)
g1 <- nrow(X)
gs <- sqrt(nrow(X))
X_gammarange = X[, which(gamma_prev == 1)]
# beta_gammaj = beta_prev[which(gamma_prev == 1)]
# err = Y - (X_gammarange * beta_gammaj)
sigma2_prev <- 1 / rgamma(1, 0.01 + n/2, 0.01 + t(err)^2)
for (i in 1:length(c_prev)) {
  cat("c_prev", c_prev, "\n")
  if (sum(as.integer(c_prev)) >= 2){
    Xj = X[, j]
    X_j = X[-j]
    X_gammarange_j = X_j[, which(gamma_prev[-j] == 1)]
    A <- t(Xj) %*% X_gammarange_j
    # update G_j based on gamma=1
    G_j <- G[-j, -j]
    G_j_gammaj = G_j[which(gamma_prev[-j] == 1), which(gamma_prev[-j] == 1)]
    # update beta_j based on gamma=1
    beta_j <- beta_prev[-j]
    beta_gammaj = beta_j[which(gamma_prev[-j] == 1)]
    B <- beta_gammaj %*% solve(G_j_gammaj)
    D <- A^t(B)
c_prev[j] <- 0
C_n<-c_prev

g_prev <- numeric()
for (i in 1:length(c_prev)) {g_prev[i]<-c_n[i]*b0*gl+(1-c_n[i])*b0 *gs}
G <- diag(g_prev)
G.gamma=G[which(gamma_prev==1),which(gamma_prev==1)]
G.gamma<-as.matrix(G.gamma)

l0<- -(n/2)*log(2*pi*sigma2_prev) -
    log(det(G.gamma))+(1/2)*log(det(t(X.gamma)%*%X.gamma))- 
    (1/(2*sigma2_prev))*(beta_prev[j]*solve(G[j,j]))*(D) +
    beta_prev[j]^2*solve(G[j,j]^2*(t(X)%*%X))[j,j] + log(0.5)

c_prev[j] <- 1
C_n<-c_prev

g_prev <- numeric()
for (i in 1:length(c_prev)) {g_prev[i]<-c_n[i]*b0*gl+(1-c_n[i])*b0 *gs}
G <- diag(g_prev)
G.gamma=G[which(gamma_prev==1),which(gamma_prev==1)]
G.gamma<-as.matrix(G.gamma)

# compute chance parameter of the conditional posterior of gamma_j (Bernoulli)
l1<- -(n/2)*log(2*pi*sigma2_prev ) -
    log(det(G.gamma))+(1/2)*log(det(t(X.gamma)%*%X.gamma))- 
    (1/(2*sigma2_prev))*(beta_prev[j]*solve(G[j,j]))*(D) +
    beta_prev[j]^2*solve(G[j,j]^2*(t(X)%*%X))[j,j] + log(0.5)
# sample gamma_j from a Bernoulli and #update
C_prev[j] <- rbinom(1, 1, exp(l1) / (exp(l1) + exp(l0)))
}

# if (sum(as.integer(c_prev)) >= 2){
#    c_new<-c_prev
#}
# return(c_new)
#update beta based on gamma=1
beta_fnc<- function(gamma_prev, X, G, sigma2_prev, Y, c_prev, beta_prev, b0){
    #cat("beta_prev", beta_prev, "\n"")
    require(MASS) ##
    require(mvtnorm)
    X_gama=X[,which(gamma_prev==1)]
    gl<- nrow(X)
    gs<- sqrt(nrow(X))
    g_prev <- numeric()
    for (i in 1:length(c_prev)) {g_prev[i]<- c_prev[i]*b0*gl+(1-c_prev[i])*b0*gs}
    G <- diag(g_prev)
    G_gama=G[which(gamma_prev==1), which(gamma_prev==1)]
    post.var.beta <- (sigma2_prev^2) * solve(t(X_gama)%%X_gama +
         solve(G_gama)%% (t(X_gama)%%X_gama)%%solve(G_gama))
    post.mean.beta <- ((1/sigma2_prev^2) * post.var.beta)%%t(X_gama)%%Y
    #cat("post.var.beta", post.var.beta, "\n")
    #update
    beta_post <- rmvnorm(n=1, mean=post.mean.beta, sigma=post.var.beta)
    for (j in 1:length(gamma_prev)){
        beta_prev[which(gamma_prev==1)]<-beta_post
        beta_prev[which(gamma_prev==0)]<-0
    }
    #return(beta_prev)
    beta_new<-beta_prev
    return(beta_new)
}

#update initial beta based on gamma=1
betgam_fnc<- function(gamma_prev, beta_prev, X){
    p <- ncol(X)
    beta_gama=beta_prev[which(gamma_prev==1)]
    return(beta_gama)
}

#update sigma2 based on gamma=1
sigma2_fnc<- function(X_gama, beta_gama, X, Y){
    n <- nrow(X)
    err=Y-(X_gama%%beta_gama)
    sigma2_prev<- 1 / rgamma(1, 0.01 + n/2, 0.01 + t(err) %% err / 2)
    sigma2_new<-sigma2_prev
    return(sigma2_new)
source("Datagen_function2.R")
source("Gibbs_Functions4.R")
set.seed(123)
out <- mydata(p=15, n=100, rho=0.6)
y <- out$Y
dim(y)
x <- out$x
dim(x)

# res is where we store the posterior samples
res <- matrix(NA, nrow = niter, ncol = 3*p + 1)
colnames(res) <- c(
  paste0('gamma', seq(p)),
  paste0('c', seq(p)),
  paste0('beta', seq(p)),
  'sigma2')

# take the MLE estimate as the values for the first sample
m <- lm(y ~ x - 1)

## initial value
res[1, seq(p)] <- rbinom(p, 1, 0.5)# initial value for gamma
gamma0_prev <- res[1, seq(p)]

res[1, seq(p+1, 2*p)] <- rbinom(p, 1, 0.5)# initial value for c
c_prev <- res[1, seq(p+1, 2*p)]

res[1, seq(2*p+1, 3*p)] <- m[1]# initial value for beta
beta0_prev <- res[1, seq(2*p+1, 3*p)]

res[1, 3*p+1] <- 3## initial value for sigma
sigma2_prev <- res[1, 3*p+1]

#-----------------------------
gl <- mrow(x)
gs <- sqrt(mrow(x))
g_prev <- numeric()
for (j in 1:length(c_prev)) {g_prev[j] <- c_prev[j]^b0*gl*(1-c_prev[j])^b0 * gs}
g <- diag(g_prev)

#-----------------------------
# we start running the Gibbs sampler
for (i in seq(2, niter)) {
    # Sampling gamma
gibbs.gamma <- gamma_fnc(gamma_prev, x, y, G, beta_prev)
    X.gamma <- xgam_fnc(gamma_prev, X)

    # Sampling c
    gibbs.c <- c_fnc(gamma_prev, X, G, beta_prev, sigma2_prev, c_prev, b0)

    G.gamma <- GGamma_fnc(X, gamma_prev, c_prev, b0)

    # Sampling Beta from multivariate Gaussian
    gibbs.beta <- beta_fnc(gamma_prev, X, G, sigma2_prev, Y, c_prev, beta_prev, b0)

    beta.gamma <- betgam_fnc(gamma_prev, beta_prev, X)

    gibbs.sigma2 <- sigma2_fnc(X.gamma, beta.gamma, X, Y)

    # Store results: add new samples
    res[i, ] <- c(gibbs.gamma, gibbs.c, gibbs.beta, gibbs.sigma2)

    # First, get all the values of the previous time point
    gamma_prev <- gibbs.gamma
c_prev <- gibbs.c
    beta_prev <- gibbs.beta
    sigma2_prev <- gibbs.sigma2
}

# remove the first nr_burnin number of samples
res[-seq(nr_burnin), ]
return(res)
## Fixed effects matrix

```r
X_mat <- function(p, n, rho){
  require(mvtnorm)
  mean.X <- rep(0, p)
  Sigma.X <- matrix(rho, p, p)
  diag(Sigma.X) <- 1

  Xpred <- rmvnorm(n, mean.X, Sigma.X)
  cat.var <- rbinom(n, 1, 0.5)
  int.var <- Xpred[,1] * Xpred[,p]
  V1.cat <- Xpred[,1]*cat.var
  X <- cbind(Xpred, cat.var, V1.cat, int.var)

  return(X)
}

#X <- X_mat(5, 30, rho=0)
```

## mydata

```r
mydata <- function(p, n, rho){
  Beta <- c(1, rep(0, p), 2, rep(0, (p-1))))
  X <- X_mat(p, n, rho)
  Error <- rnorm(n, 0, 1)
  Y <- X%%Beta + Error
  Out=list(Y=Y, X=X)
  return(Out)
}

#Y <- mydata(p=5, n=30, rho=0)
```
## Fixed effects matrix

X_mat <- function(p, n, rho){
  require(mvtnorm)
  mean.X <- rep(0, p)
  Sigma.X <- matrix(rho, p, p)
  diag(Sigma.X) <- 1

  xpred <- rmvnorm(n, mean.X, Sigma.X)
  cat.var <- rbinom(n, 1, 0.5)
  int.var <- xpred[,1] * xpred[,2]
  vl.cat <- xpred[,1] * cat.var
  X <- cbind(xpred, cat.var, vl.cat, int.var)

  return(X)
}

X <- X_mat(15, 30, rho=0.6)

##

mydata <- function(p, n, rho){
  Beta <- c(1, 1.5, 2, 1.8, rep(0, p-4), 0, rep(0, (p-1)));
  X <- X_mat(p, n, rho)
  Error <- rnorm(n, 0, 1)
  Y <- X*%*%Beta + Error
  Out=list(Y=Y, X=X)
  return(out)
}

Y <- mydata(p=15, n=30, rho=0.6)
Appendix A

References


