ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

Fast Bayesian feature selection for high-dimensional data using mixtures of g-priors

Author:
Koroniadis Konstantinos

Supervisors:
Parolli Roberta
(Università Cattolica del Sacro Cuore)
Ntzoufras Ioannis
(AUEB)

M.Sc. Thesis
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
September 2023

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Fast Bayesian feature selection for high-dimensional data using mixtures of g-priors

Συγγραφέας:
Κορωνιάδης Κωνσταντίνος

Επιβλέποντες:
Parolli Roberta
Ντζούφρας Ιωάννης

Μεταπτυχιακή Διατριβή
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα, Ελλάδα
Σεπτέμβριος 2023

ACKNOWLEDGEMENTS

This is the revised version of the thesis after addressing for the comments of the examining committee. Any remaining typos and notational conflicts is due to the author of this thesis.

I'd like to extend my heartfelt gratitude to Dr. Ioannis Ntzoufras for his invaluable guidance and unwavering patience throughout this journey. Additionally, I extend my sincere appreciation to Roberta Paroli, who not only served as a supervisor but also played a pivotal role in steering this thesis towards fruition.

I am deeply thankful to those individuals whose support has taken diverse forms, each contributing uniquely. Foremost, my thesis supervisors deserve special mention for their constant availability, constructive feedback, and patient responses to my inquiries. Roberta Paroli's contributions, in particular, have been indispensable.

Acknowledgment is also due to the Erasmus program and the collaborative efforts of both institutions, which have enriched the foundation of this work. Among these expressions of gratitude, I can't overlook the immense support from my family, whose unwavering belief in my choices has been a driving force.

Lastly, I reserve a profound thank you for my girlfriend Φωτεινή Σ, a constant source of strength and encouragement, whose unwavering presence has been pivotal in sustaining my determination to persist.

ABSTRACT

This master's thesis delves into the realm of Bayesian model selection and variable inclusion techniques within the context of linear models. The journey commences with an establishment of the normal linear model, followed by an introduction to the Bayesian framework. This serves as the foundation for a comprehensive exploration of various techniques that shape the landscape of modern statistical analysis.

The investigation unfolds with an exploration of the g-prior framework, encompassing the broader realm of Bayesian model selection. This journey encompasses many Bayesian techniques, the exploration then veers into the Stochastic Selection methods.

Leveraging the groundwork laid by Liang et al. (2008), this thesis probes paradoxes and modifications of the g-prior, particularly in the context of mixture models. The culmination of this exploration results in the development of the FBVS algorithm.Additionally, this thesis presents a compelling initial toy example that illustrates the various choices of the g-prior and mixtures of g-prior based on Liang's work.

A detailed analysis of the FBVS algorithm reveals its adaptability across various scenarios. The depiction of correlation thresholds adds practical value to the algorithm's application. Furthermore, the algorithm's extension to mixtures of the g-prior showcases innovative problem-solving and resilience.

Empirical validation through simulations reaffirms the algorithm's utility. The thesis concludes by affirming the significance of these methodologies and their contributions to the broader field of Bayesian methods. The combinations of parameter alpha, sample size, and possible features are thoroughly examined, demonstrating that as the dataset expands, the accuracy of the methodology also increases.

In summary, this thesis not only unravels the intricate fabric of Bayesian model selection but also underscores its pivotal role in contemporary statistical analysis. The methodologies and insights presented here are poised to shape the trajectory of model selection, variable inclusion, and Bayesian inference.

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία εξετάζει τον κόσμο της επιλογής μοντέλου βασιζόμενη σε Μπευζιανές τεχνικές και της περιλαμβανομένης μεταβλητής μέσα στο πλαίσιο των γραμμικών μοντέλων. Ο αφηγηματικός ταξιδιωτικός δρόμος αρχίζει με την εγκαθίδρυση του κανονικού γραμμικού μοντέλου, ακολουθούμενο από μια εισαγωγή στο Μπευζιανό πλαίσιο. Αυτό λειτουργεί ως το θεμέλιο για μια εκτενή εξερεύνηση διαφόρων τεχνικών που διαμορφώνουν το τοπίο της σύγχρονης στατιστικής ανάλυσης.

Η έρευνα αναπτύσσεται με μια εξερεύνηση του πλαισίου της $g - prior$, περιλαμβάνοντας τον ευρύτερο κόσμο της επιλογής μοντέλου βασισμένης σε Μπευζιανές τεχνικές. Αυτό το ταξίδι περιλαμβάνει πολλές Μπευζιανές τεχνικές, η εξερεύνηση στρέφεται στις μεθόδους Στοχαστικής επιλογής.

Αξιοποιώντας το θεμέλιο που έθεσε η *Liang* και συνεργάτες (2008), αυτή η διατριβή εξετάζει παράδοξα και τροποποιήσεις της $g - prior$, ιδιαίτερα στο πλαίσιο των μοντέλων μείξεων. Η κορύφωση αυτής της εξερεύνησης οδηγεί στην ανάπτυξη του αλγορίθμου *FBVS*. Επιπλέον, αυτή η διατριβή παρουσιάζει ένα συναρπαστικό πρώτο παράδειγμα, που επιδεικνύει τις διάφορες επιλογές της $g - prior$ και των μείξεων της $g - prior$ βασισμένες στο έργο της *Liang*.

Μια λεπτομερής ανάλυση του αλγορίθμου *FBVS* αποκαλύπτει την προσαρμοστικότητά του σε διάφορα σενάρια. Η απεικόνιση των ορίων συσχέτισης προσθέτει πρακτική αξία στην εφαρμογή του αλγορίθμου. Επιπλέον, η επέκταση του αλγορίθμου σε μείξεις της $g - prior$ αναδεικνύει την καινοτομία και την ανθεκτικότητα στην επίλυση προβλημάτων.

Η εμπειρική επικύρωση μέσω προσομοιώσεων επαναβεβαιώνει τη χρησιμότητα του αλγορίθμου. Η διατριβή καταλήγει επιβεβαιώνοντας τη σημασία αυτών των μεθοδολογιών και της συμβολής τους στον ευρύτερο τομέα των Μπευζιανών μεθόδων. Οι συνδυασμοί των παραμέτρων αλφα, μεγέθους δείγματος και δυνητικών μεταβλητών, εξετάζονται λεπτομερώς, καταδεικνύοντας ότι με την αύξηση του συνόλου δεδομένων, η ακρίβεια της μεθοδολογίας αυξάνεται επίσης.

Συνολικά, αυτή η διατριβή δεν απλώς αποκαλύπτει το περίπλοκο ύφασμα της επιλογής μοντέλου βασισμένης σε Μπευζιανές τεχνικές, αλλά υπογραμμίζει επίσης τον ζωτικό της ρόλο στη σύγχρονη στατιστική ανάλυση. Οι μεθοδολογίες και οι προτάσεις που παρουσιάζονται εδώ έχουν τη δυνατότητα να διαμορφώσουν την πορεία της επιλογής μοντέλου, της περιλαμβανομένης μεταβλητής και της βαϊασιανής είσοδου.

Contents

List of Tables

List of Figures

Table 1: Notation Used in the Thesis

Symbol	**Description**
Y	Dependent variable
X_1	Independent variables
μ	Mean
σ^2	Variance
β_0	Intercept
β_p	Coefficients
ϵ	Error term
θ	Model parameters
y	Data
X	Design matrix
τ^2	Fixed hyperparameter controlling prior scale
a_0	Fixed hyperparameter controlling prior shape
b_0	Fixed hyperparameter controlling prior scale
λ	Hyperparameter controlling regularization
γ	Binary variables
M	Model M
h	Level of regularization
p	Number of predictor variables
κ	Complexity parameter
$P(M)$	Prior probability of model M
P	Total number of potential predictor variables
p_M	Number of predictor variables in model M
β_{MAV}	Model-averaged regression coefficients
β_M	Coefficients from model M
η	Linear predictor
Σ	Variance-covariance matrix
k_j	Hyperparameter
g	Controls shrinkage or regularization on coefficients
ϕ	Precision parameter
X^T	Transpose of matrix X
β_a	Prior mean
BF	Bayes Factor
M_N	Null model
M_F	Full model
${R_\gamma}^2$	Coefficient of determination of model M_γ
Γ	Gamma function
$X_{-\gamma}$	Columns of design matrix ignored by model M_γ
F_γ	F statistic
n	Sample size
$\hat{a}$	Ordinary least squares estimate of intercept parameter
${}_2F_1$	Gaussian hypergeometric function
L	Maximum likelihood

Chapter 1

Introduction

1.1 Normal Linear Models

Normal linear models, also known as linear regression models, describe the relationship between a continuous dependent variable Y and one or more independent variables $X_1, X_2, \ldots, X_p$. In these models, it is assumed that Y follows a normal or Gaussian distribution with mean μ and variance σ^2:

$$Y \sim \text{Normal}(\mu, \sigma^2) \tag{1.1}$$

The relationship between Y and $X_1, X_2, \ldots, X_p$ is assumed to be linear, with the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{1.2}$$

where β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients for the independent variables, and ϵ, where $\epsilon \sim Normal(0, \sigma^2)$ is the error term that captures the random variation in Y not explained by the independent variables.

Estimating the values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ that provide the best-fit line or equation that represents the relation between the variables is the aim of normal linear models. This is accomplished by reducing the sum of squared errors between the observed Y values and the model-predicted values. The method of least squares is used to estimate the parameters in order to identify the values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ that minimize the sum of squared errors.

Normal linear models find extensive application across numerous domains, such as economics, psychology, and social sciences, for the purpose of predicting and comprehending the determining factors that impact the response variable. With their versatility and strength, normal linear models offer a valuable and adaptable mechanism for scrutinizing data and forecasting outcomes, provided the conditions and assumptions of the model are satisfied.

1.1.1 Bayesian Inference

Bayesian inference is another field where normal models can be applied. In Bayesian analysis, probability distributions are used to represent uncertainty about the model parameters, while incorporating prior knowledge or beliefs about them, followed by updating them based on the observed data using Bayes' theorem. For normal linear models, Bayesian inference can facilitate the estimation of the posterior distribution of the parameters, considering the data and prior information.

In a Bayesian normal linear model, the prior distribution for the model parameters θ (which can include the intercept, coefficients, and variance) is specified as $p(\theta)$. The likelihood of the data y given the parameters is given by the normal distribution(N):

$$p(y|\,\theta) = \prod_{i=1}^{n} Normal(y_i|x_i^\top \beta, \sigma^2) \tag{1.3}$$

where x_i is the i-th row of the design matrix X, β is the vector of coefficients, and σ^2 is the variance of the error term. The posterior distribution of the parameters given the data is then obtained using Bayes' theorem:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{1.4}$$

To get samples from the posterior distribution, Markov chain Monte Carlo (MCMC) techniques like Gibbs sampling or Metropolis-Hastings can be utilized. After obtaining samples from the posterior, it is possible to construct posterior means, credible intervals, and other posterior summaries.

In comparison to conventional frequentist approaches, Bayesian inference has a number of advantages, including the capacity to take into account previous knowledge, the capacity to quantify uncertainty in the estimates, and the capacity to generate predictions using posterior predictive distributions. Prior distributions must also be specified for Bayesian inference, although they can be arbitrary and have an impact on the outcome. Bayesian inference can also be computationally demanding, particularly for complicated models or huge datasets.

In general, Bayesian inference offers a strong and adaptable framework for examining normal linear models, particularly where prior information and knowledge or uncertainty are significant factors. However, while utilizing Bayesian inference, it is important to pay close attention to the prior distributions that are chosen as well as the computing complexity of the approaches.

1.1.2 Normal Linear Models with Standard Priors

Standard priors for the parameters, such as a normal prior for the coefficients and an inverse-gamma prior for the variance, are frequently used in Bayesian inference for normal linear models. These priors are referred regarded be "standard" because they are frequently used as the default option in software programs and mainly because they have straightforward and well-known mathematical properties.

A common choice for the prior distribution of the coefficients is a normal distribution with mean zero and a fixed variance, such as:

$$\beta \sim Normal(0, \tau^2 I) \tag{1.5}$$

where τ^2 is a fixed hyperparameter that controls the scale of the prior. This prior assumes that the coefficients are centered around zero and have similar variances.

For the variance of the error term, a common choice is the inverse-gamma distribution:

$$\sigma^2 \sim \text{Inverse-Gamma}(a_0, b_0) \tag{1.6}$$

where a_0 and b_0 are fixed hyperparameters that control the shape and scale of the prior. This prior assumes that the variance of the error term is positive and has a non-zero probability at zero, but allows for a wide range of possible values.

Using these standard priors can make more simple the Bayesian analysis process and make the outcomes more consistent across various datasets and models. Nevertheless, it should be emphasized that the selection of priors can substantially affect the posterior distribution and, as a result, the inference that follows. In certain situations, it may be suitable to utilize more explicit or informed priors that are rooted in prior knowledge or expert input.

1.1.3 Normal Linear Models with g Priors

Normal Linear Models that feature g-priors belong to the category of Bayesian linear regression models, which incorporate a distinctive form of informative prior distribution for the regression coefficients. The underlying concept behind g-priors is to facilitate a selection of the degree of shrinkage or regularization applied to the coefficients, guided by the data itself.

In a normal linear model with g-priors, the prior distribution for the regression coefficients is given by:

$$\beta \sim Normal(0, g(\lambda)(X^\top X)^{-1}) \tag{1.7}$$

where λ is a hyperparameter that controls the amount of regularization or shrinkage, and $g(\lambda)$ is a function that depends on λ and is chosen to satisfy certain desirable properties.

The g-prior is constructed to strike a balance between two objectives: reducing the coefficients to zero to decrease overfitting and increase generalization, and allowing the coefficients to vary when the data strongly dismisses the null hypothesis that they are zero. The selection of λ, which affects the regularization's strength, can be made using a variety of techniques, including cross-validation or Bayesian model selection criteria.

G-priors give a data-driven approach to choosing the level of regularization and can be more robust and flexible in handling diverse types of data and models than conventional ridge or lasso regression approaches that use fixed penalty parameters. Nevertheless , the selection of the hyperparameter λ and the function $g(\lambda)$, as well as the computational cost of the Bayesian inference, must be carefully taken into account in the g-prior technique.

1.2 Bayesian Variable Selection Problem

A statistical method called Bayesian variable selection is used to define the subset of variables that are most significant to a particular problem. The fundamental idea underlying this method is to estimate, given the data and a prior distribution across the set of potential models, the likelihood that each variable is meaningful using Bayesian inference.

In order to define formally the problem , consider that we have a set of data points $y = y_1, y_2, \ldots, y_n$ and a set of p predictor variables $X = X_1, X_2, \ldots, X_p$. Our objective is to select a subset of variables $S \subseteq X$ that are most relevant to predicting Y. We can represent the problem as a set of binary variables $\gamma = \gamma_1, \gamma_2, \ldots, \gamma_p$, where $\gamma_i = 1$ if X_i is included in the subset S and $\gamma_i = 0$ otherwise.

The Bayesian approach to variable selection involves specifying a prior distribution over the space of possible models, which is typically done by assigning a prior probability to each possible subset of variables. A common choice is the so-called "spike-and-slab" prior, which places a spike at the null model with no predictors and a slab over the remaining models, with a relatively flat distribution over the space of possible subsets. The approach was further significantly developed by Madigan & Raftery (1994)(22) and George & McCulloch (1997)(15).

1.2.1 Model specification

Each model M_γ is associated with binary indicator vector $\gamma = [0,1]^p$ where $\gamma_i = 1$ means that the variable X_i is included in the model M_γ and $\gamma_i = 0$ that is excluded.
In the classic linear regression model, we consider a P-dimensional vector γ_p, which is drawn from a Bernoulli distribution with a prior inclusion probability denoted as h. This applies for all values of p, ranging from 1 to P.

$$\gamma_p \sim \text{Bernoulli}(h) \quad \text{for } p = 1, 2, \ldots, P \tag{1.8}$$

To give an example , if we have 1000 covariates and we suppose h is 0.01 then the expected number of a priori covariance that we will include in the model are $h * P$, in this case 10. This does not mean that we look the models only that have exactly 10 covariates , this is just explain how h is being interpreted (the level of regularization). A Bernoulli prior distribution is often used as the default choice when there is no specific prior information available about the initial probabilities of including specific variables.
Moving on to the normal likelihood, we have:

$$Y_n \sim \text{Normal}(\beta_0 + X_{n,\gamma}\beta_\gamma, \sigma^2) \tag{1.9}$$

Here, Y_n represents the response variable, and $\beta_0 + X_{n,\gamma}$ is the linear predictor.

The covariance term σ^2 in the normal likelihood has a prior distribution, commonly modeled as an inverse gamma distribution:

$$\sigma^2 \sim \text{Inverse-Gamma}(a_0, b_0) \tag{1.10}$$

where a_0 and b_0 are fixed hyperparameters that control the shape and scale of the prior. This prior allows for a wide range of possible values. To be more specific about the hyperparameters :

- α_0 controls the scale of the distribution. A higher value of α_0 makes the distribution more peaked around its mean, indicating stronger prior beliefs about the variance.
- β_0 is related to the shape of the distribution. A higher value of β_0 makes the distribution more spread out, indicating less certainty about the variance.

Moving on to the coefficients, we have:

The intercept term, β_0, which follows a normal distribution:

$$\beta_0 \sim \text{Normal}(0, \sigma^2 \tau_{\text{intercept}}^{-1}) \tag{1.11}$$

The coefficients β_γ, where the subscript γ signifies their association with the set of covariates being considered. These coefficients also follow a normal distribution:

$$\beta_\gamma \sim \text{Normal}(0, \sigma^2 \tau_\gamma^{-1}) \tag{1.12}$$

The subscript γ in β_γ indicates that the number of covariates entering the model is determined by the vector γ. For example, if $\gamma = (1, 1, 0, 0, \ldots, 0)$, then β_γ is two-dimensional, reflecting the inclusion of only two covariates.

These equations collectively define the Bayesian variable selection for linear regression model presented in this introduction.

1.2.2 Spike and slab prior

The idea is to introduce a latent variable $\gamma_j = 0$ or 1, the normal mixture by

$$\beta_j|\gamma_j \sim (1-\gamma_j)Normal(0, {\tau_j}^2) + \gamma_j Normal(0, {c_j}^2{\tau_j}^2) \tag{1.13}$$

It is a mixture of two Normal distribution that one is a large variance slab and one is low variance spike. When γ_j=0, $\beta \sim Normal(0, {\tau_j}^2)$ and when $\gamma_j = 1 \sim Normal(0, {c_j}^2{\tau_j}^2)$ we need to set τ small so, the β_j would be 0. Also about the c_i is the opposite choice, must be over 1 and a large number for the case when we have to include the variable X_j the β_j would not be 0. The $\gamma_j \sim Bernoulli(1/2)$ with usual choice of probability of success equals to $1/2$, this is known as indifference, or uniform prior, George and McCulloch (1993)(13). The idea behind this specification is to allow the regression weight β_j to be exactly zero.

1.3 Variable Selection Methods

1.3.1 Bayesian Model Average

Bayesian Model Averaging (BMA)(16) is a statistical approach that considers uncertainty in model selection when estimating a linear regression model. BMA involves combining predictions from multiple models, where each model corresponds to a different combination of predictor variables. The combined predictions are weighted by their posterior model probabilities, resulting in a more accurate estimate of the relationship between predictor variables and the response variable.

A prior probability distribution is given to each potential model in order to calculate the posterior model probabilities. Following that, using Bayes' theorem—which states that the probability of a model given the data is proportional to the product of the prior probability of the model and the likelihood of the data given the model—the posterior probabilities are updated based on the available data. The likelihood of the data given a model M and model parameters θ is calculated using the normal probability density function:

$$P(y|M,\theta) = Normal(Y|X\beta,\sigma^2) \tag{1.14}$$

where N is the normal distribution, X is the design matrix of predictor variables, β is the vector of regression coefficients, and σ^2 is the variance of the residuals.

The prior probability of a model M is typically calculated as a function of the number of predictor variables p included in the model and a complexity parameter κ:
To give more details , a prior probability of a model M can be specified as :

$$P(M) = \prod_{i=1}^{p} \kappa_i^{\gamma_i}(1-\kappa_i)^{1-\gamma_i} \tag{1.15}$$

The choice of κ can impact the results of BMA. Different values of κ can lead to different model selection outcomes and posterior probabilities.
Let's consider κ_i as the prior probability that β_i is not equal to zero within a regression model. When $\kappa_i = 1$, it provides strong evidence that the corresponding variable should indeed be included in the model M. A common and straightforward approach is to assign the same value of k to all variables. For instance, if we have strong prior knowledge that most variables are likely irrelevant, we might choose a small κ (close to 0). By setting κ_i to 0.5 for all variables, we establish a uniform prior across the model space. If we believe that most variables are relevant, we might choose a larger κ (close to 1).
For the upcoming example, we will assume a common value for κ.
So for example if we have in total 5 variables (P= 5) and in a specific model M only the first three are chosen ($\gamma = (1,1,1,0,0)$) , then the prior probability of the model M can calculated as :

$$P(M) = \prod_{i=1}^{5} \kappa^{\gamma_i}(1-\kappa)^{1-\gamma_i} = \kappa\kappa\kappa(1-\kappa)(1-\kappa) = \kappa^3(1-\kappa)^2$$

As a result the final equaltion is the following :

$$P(M) = \kappa^{p_M}(1-\kappa)^{P-p_M} \tag{1.16}$$

In this equation:

- $P(M)$ represents the prior probability of the model M.
- P is the total number of potential predictor variables.
- p_M is the number of predictor variables included in the model M.
- κ is a hyperparameter that determines the prior probability of including a predictor variable in the model.

After computing the posterior model probabilities, the predictions are generated by taking the weighted average of the predictions of all the models (weighted by their posterior probabilities). This approach provides a more reliable and consistent estimate of the relationship between the predictor variables and the response variable, since it considers the uncertainty in model selection. The model-averaged regression coefficients can be calculated using the following formula:

$$\beta_{MAV} = \sum_{M} P(M|y)\beta_M \tag{1.17}$$

where β_{MAV} is the vector of model-averaged regression coefficients, β_M is the vector of coefficients from model M, and the sum is taken over all possible models.

BMA offers a more complete and precise understanding of the relationship between the response variable and predictor variables, especially in scenarios where the number of potential predictor variables is high or the relationship between the variables is complex and challenging to model. By incorporating the uncertainty in model selection directly, BMA addresses the constraints of conventional linear regression analysis and delivers more dependable estimates of the regression coefficients.

1.3.2 Bayesian Adaptive Sampling (BAS)

Bayesian adaptive sampling is another powerful method used in linear models to determine the best model from a set of candidate models. The essential concept is explained in Clyde et al. ;(6) MCMC algorithms are made to sample from finite model spaces with replacement. Subsequently, every model is a posteriori ranked or chosen as the model with the highest probability by counting the visits on each model, or, more specifically, by calculating the MCMC model frequencies. Thus, resampling over model space doesn't really yield any new information, and sampling over model spaces without replacement might yield a more effective model-search method.

One of the key components of this method is the prior distributions used on the coefficients. Zellner's g prior is a popular choice for the prior distribution, and it is based on a mixture of g-priors including Zellner and Siow prior, hyper-g prior, hyper-g/n prior, local and global empirical g-prior, as well as criteria such as AIC and BIC.

The main objective of Bayesian adaptive sampling in linear regression is to find the most suitable model with the highest posterior probability. This is achieved by iteratively sampling from the posterior distribution and evaluating the posterior probability of each model. The model with the highest probability is then selected and the process is repeated until convergence is achieved. To be more specific , BAS samples near the median probability model, providing that the sampling probabilities are the marginal inclusion probabilities. This means that it is able to identify the model with the highest posterior probability, while also providing information about the uncertainty associated with the model selection process.

Advantage

One of the challenges in Bayesian model selection is estimating the marginal inclusion probabilities of the covariates, which are not known beforehand. BAS addresses this issue by estimating the marginal inclusion probabilities adaptively during the sampling process. This allows for accurate and efficient model selection, even when the number of covariates is approximately 30 can sample all over the model space.

Another one of the strengths of BAS is that it provides perfect samples when the number of covariates is larger and sampling is unavoidable, under the condition of orthogonality or of limiting dependence. This is important because it allows for accurate and reliable estimation of the posterior distribution of the models. As a result,BAS is a useful tool for analyzing complex data sets.

1.3.3 Stochastic Search Variable Selection (SSVS)

This method, called Stochastic Search Variable Selection *SSVS* , was introduced by George and McCulloch (14). The method works by treating variable inclusion as a Bernoulli trial and using a mixed prior to model the possible presence or absence of each variable. The prior distribution is chosen such that it is relatively non-informative, and the method uses a Markov Chain Monte Carlo (MCMC) algorithm to generate posterior samples of the model. To update the model, the algorithm proposes a new state by randomly adding or removing a variable. The decision to accept or reject this proposal is based on the posterior probability of the new state. It was originally applied in linear models (such as Normal Linear Model) and after that it expanded to a wide variety of other different models such as Generalized Linear Models (George et al.)(13) , multivariate normal regression (Brown et al.)(5) , and log linear models for multi-way contingency tables (Ntzoufras et al.)(25).

More analytically , suppose we have all possible models models and p covariates. We consider that m is the model indicator and that β_j are the parameters involved in the linear predictor of a GLM. We consider also a binary inclusion indicator γ_j for j = 0 , 1 , . . . P where :

- γ_j is equal to 1 when X_j is including into the model
- γ_j is equal to 0 when X_j is not included into the model.

The key feature of this method and at the same time its main advantage is the fact that it maintains a constant parameter space across all models. That is, more detailed, when $\gamma_j = 0$ (so the corresponding variable should be excluded from the model) then the posterior of β_j accepts values very close to zero and as a result β_j does not completely eliminate. In order to achieve that, when this situation occurs the idea is to set a very strong prior centered to zero. So the linear predictor for all the models is :

$$\eta = \sum_{j=1}^{p} X_j \beta_j = X\beta \tag{1.18}$$

To implement the above described algorithm a mixture prior of two Normal distributions was proposed to be used as a prior for each β_j given the indicator γ_j.

$$\beta_j | \gamma_j \sim \gamma_j Normal(0, \Sigma_j) + (1 - \gamma_j) Normal(0, k_j^{-2} \Sigma_j) \tag{1.19}$$

Where :

The Σ is the prior variance-covariance matrix Σ_j and the value of the parameter k_j is set equal to a large number (>1) , so in case of γ_j =1 , then according to the prior the impact of the variable β_j would be non-zero, hence it would be a candidate to be included as a regressor and the posterior distribution will primarily depend on the data.

As a result, the following types give the whole conditional posterior distributions of β_j and γ_j using a simple Gibbs sampler algorithm.

Update β_j from :

$$f(\beta_j | y, \gamma, \beta_{-j}) \propto f(y | \beta, \gamma) \times f(\beta_j | \gamma_j) \tag{1.20}$$

Update the γ_J from a Bernoulli distribution with success probability $O_j/(1 + O_J)$:

$$O_j = \frac{f(\gamma_j = 1 | y, \gamma_{-j}, \beta)}{f(\gamma_j = 0 | y, \gamma_{-j}, \beta)} = \frac{f(\beta | \gamma_j = 1, \gamma_{-j})}{f(\beta | \gamma_j = 0, \gamma_{-j})} \times \frac{f(\gamma_j = 1, \gamma_{-j})}{f(\gamma_j = 0, \gamma_{-j})} \tag{1.21}$$

Where γ_{-j} include all possible γ except from the one related to the specific j. Also the first

fraction is prior ratio for the parameteres and the other is the prior model odds. Except from the main advantage of this algorithm, which as mentioned above is the constant dimension space for all models, SSVS method also provides the same likelihood for all models, leads to the convergence of the algorithm and it is simple to implement for any GLM.
On the other hand the results are not very similar to usual variable selections due to presence of parameters even when we have γ_J equal to zero, also the choice for k_J is difficult, and the independent priors may occuse problem when we have collinearity. After this method, a number of Gibbs sampling-based techniques with a similar approach appeared , including the Kuo and Mallick sampler.

1.3.4 Kuo & Mallick (KM) Sampler

A different approach for incorporating the indicator variable into Bayesian variable selection involves integrating γ directly into the likelihood equation. The following technique has two points that differentiate it from SSVS method. First of all the model have distinct model structures compared to SSVS (where the likelihood relies on the indicator). In adittion , it differ in the process of selection of prior specifications.

The Kuo-Mallick(18) algorithm is a different MCMC approach to Bayesian variable selection that is capable of selecting variables in regression models. This algorithm relies on the spike-and-slab prior, which , as mentioned above, is a well-known prior utilized in Bayesian variable selection.

The method uses a simple Gibbs sampler and is ease to adopt for any GLM with reasonable results. Likelihood depends on γ_j via the following linear predictor :

$$\eta = \sum_{j=1}^{p} \gamma_j \, X_j \beta_j \tag{1.22}$$

Where the X_j is a matrix with elements X that corresponds to the coefficient β_j of the jth feature. The K-M method assumes that the indicators and effects are independent a priori, so independent priors are placed on each γ_j and β_j such that :

$$f(\beta, \gamma) = f(\beta)f(\gamma) = f(\beta|\beta_{-\gamma})f(\beta_{-\gamma})f(\gamma) \tag{1.23}$$

The actual prior is :

$$f(\beta_\gamma) = \int f(\beta, \beta_{-\gamma})d\beta_{-\gamma} \tag{1.24}$$

and the corresponding posterior required for the Gibbs algorithm is the following:
When $\gamma_j = 1$

$$f(\beta_j|y, \gamma, \beta_{-j}) \propto f(y|\beta, \gamma)f(\beta_j|\beta_{-j}) \tag{1.25}$$

when $\gamma_j = 0$

$$f(\beta_j|y, \gamma, \beta_{-j}) \propto f(\beta_j|\beta_{-j}) \tag{1.26}$$

The indicator γ_j is updating using the Bernoulli distribution with probability of success:

$$p = \frac{O_j}{1 + O_j} \tag{1.27}$$

$$O_j = \frac{f(\gamma_j = 1|y, \gamma_{-j}, \beta)}{f(\gamma_j = 0|y, \gamma_{-j}, \beta)} = \frac{f(y|\gamma_j = 1, \gamma_{-j}, \beta)}{f(y|\gamma_j = 0, \gamma_{-j}, \beta)} \times \frac{f(\gamma_j = 1, \gamma_{-j})}{f(\gamma_j = 0, \gamma_{-j})} \tag{1.28}$$

As in the previous method the 1st fraction is the likelihood ratio and the second is again the prior model odds. Thus, variable selection in the K-M method is essentially a discrete process where predictors are either excluded or included in the model at every iteration. The MCMC algorithm to fit the model does not require any tuning. However, as pointed out by

O'Hara and Sillanpaa (2009)(27), when γ_j = 0, mixing might be poor if the prior on the β_j is too vague.

Prior is specified only for the full model, but this simplicity may also be a drawback, as there is no flexibility here to alter the method to improve efficiency. The collinearity is still a problem, selection of the prior for the full model may result to strange priors for each model. Another advantage of this method is that doesn't need to apply all models as SSVS does.

1.3.5 Gibbs Variable Selection (GVS)

This method is also uses an indicator variable as part of the model equation and was introduced by Dellaportas et al. (2002)(9). First of all we need to specify the prior distribution for (β, γ) with the underline hierarchical structure $f(\gamma, \beta) = f(\gamma) f(\beta|\gamma)$. Splitting β into (β_γ ,$\beta_{-\gamma}$), the prior f($\beta|\gamma$) can be split from the model prior into $f(\beta_\gamma|\gamma)$ and the pseudoprior $f(\beta_{-\gamma}|\beta_\gamma, \gamma)$ which is not affect the posterior since the $\beta_{-\gamma}$ is independent of the likelihood. The full conditional posterior distributions for the features are given by:

$$f(\beta_\gamma|\beta_{-\gamma}, \gamma, y) \propto f(y|\beta, \gamma) f(\beta_\gamma|\gamma) f(\beta_{-\gamma}|\beta_\gamma, \gamma) \tag{1.29}$$

$$f(\beta_{-\gamma}|\beta_\gamma, \gamma, y) \propto f(\beta_{-\gamma}|\beta_\gamma, \gamma) \tag{1.30}$$

and for the variable indicator γ_j is updating by the Bernoulli distribution via:

$$\gamma_j|\beta, \gamma_{-j}, y \; Bernoulli \frac{O_j}{1 + O_j} \tag{1.31}$$

$$O_j = \frac{f(\gamma_j = 1|\gamma_{-j}, \beta, y)}{f(\gamma_j = 0|\gamma_{-j}, \beta, y)} = \frac{f(y|\beta, \gamma_j = 1, \gamma_{-j})}{f(y|\beta, \gamma_j = 0, \gamma_{-j})} \frac{f(\beta|\gamma_j = 1, \gamma_{-j})}{f(\beta|\gamma_j = 0, \gamma_{-j})} \frac{f(\gamma_j = 1, \gamma_{-j})}{f(\gamma_j = 0, \gamma_{-j})} \tag{1.32}$$

In this point the variable selection step depends on both Likelihood ratio as in $K\&M$ sampler and Prior ratio as in SSVS. One distinct point of this approach is that the pseudoprior $f(\beta_{-\gamma}|\beta_\gamma, \gamma)$ also affects the full conditional posterior distribution of the active model parameters β_γ. As mentioned, the full conditional posterior's dependence on the pseudoprior can be helpful when collinearity between potential variables is found. This situation can be avoided by assuming that β_γ and $\beta_{-\gamma}$ are a priori independent

So, the above equations are transforming to

$$f(\beta_\gamma|\beta_{-\gamma}, \gamma, y) \propto f(y|\beta, \gamma) f(\beta_\gamma|\gamma) \tag{1.33}$$

$$f(\beta_{-\gamma}|\beta_\gamma, \gamma, y) \propto f(\beta_{-\gamma}|\gamma) \tag{1.34}$$

By assuming prior conditional independence for all parameters given the model, Dellaportas et al. (2000(8), 2002(9)) and Ntzoufras(2009)(24) in section 11.5.2 suggest ways to further simplify the approach. They are rather restrictive, as mentioned in Ntzoufras (2009)(24), but they might be logical, for example, when the candidate variables are centered, standardized, or orthogonal. As it obvious the GVS is a natural hybrid of KM and SSVS so can combine the advantages of these methods, it is flexible since the pseudopriors can be defined in our ease for the efficiency of the algorithm. The GVS algorithm also combines the disadvantages like the problem with collinearity which is occurred in all methods and the specification of the priors which can be a drawback.

1.3.6 Comparison of Gibbs sampling variable selection

The differences of these three methods is based on the conditional probabilities in (1.21) for SSVS and (1.28) for KM sampled and last for GVS (1.32). The first ratio in (1.32) is missing in the SSVS and this happens because the $f(y|\beta, \gamma)$ in independent of γ. Also in the SSVS the parameter priors affect the posterior and as a result the densities cannot be thinking as linking densities. In the simplest method of KM sampler, which is known as "unconditional prior approach" the second ratio of (1.32) is missing in the KM as β and γ are a priori independent. Also is the simplest method because requires only the usual specification of the parameters of the full model. In the following table it is presented these differences in the ratio of conditional probability.

Table 1.1: Table with Methods and Ratios

Method	Linear predictor	PSR_j	LR_j	PR_j
GVS	Xβ	✓	✓	✓
SSVS	$\Sigma\gamma_j X_j\beta_j$			✓
KM	$\Sigma\gamma_j X_j\beta_j$		✓	

1.4 Conclusion

In conclusion, the exploration undertaken delves into the realm of Normal Linear Models, their significance in understanding relationships between dependent and independent variables, and their role as a fundamental framework for quantifying these relationships to provide insights into factors influencing the response variable.

The method of least squares stands out as a cornerstone of normal linear models, enabling the estimation of model parameters through the minimization of the sum of squared errors. This optimization technique, rooted in the principles of statistical inference, establishes a robust foundation for parameter estimation and model fitting.

The journey extended into the realm of Bayesian Inference, uncovering a powerful approach to modeling uncertainty and incorporating prior knowledge. Embracing probability distributions grants the capacity to quantify uncertainty, make predictions, and refine the understanding of underlying processes driving the data.

The exploration also encompassed the world of Bayesian Variable Selection, revealing advanced methods crafted to address high-dimensional data and complex model selection problems. Techniques such as Bayesian Model Averaging, Stochastic Search Variable Selection, Kuo & Mallick Sampler, and Gibbs Variable Selection showcase the versatility and adaptability of Bayesian approaches in handling intricate scenarios.

In culmination, the comprehension achieved transcends theoretical constructs to practical instruments applicable across diverse domains. From economics to social sciences, from predictive modeling to decision-making, the traversed concepts possess immense potential to guide through data-driven challenges.

This juncture, marking the closure of this chapter, signifies not an end, but a pivotal transition to subsequent sections of the master's thesis. Armed with this foundation, further exploration into applications, case studies, and cutting-edge advancements in statistical modeling and data analysis unfolds.

1.5 Thesis Outline

As a result of all the above, this thesis aims to investigate and analyze fast Bayesian feature selection for high-dimensional data using mixtures of g-priors.
This will be accomplished through a review of existing literature, an examination of relevant statistical methodologies, simulations to validate proposed techniques, and the assessment of both toy examples and real-world case studies.

Preview of thesis structure

The thesis has been divided into four chapters including this one which is the introduction, the presentation of Bayesian variable selection problem and the presentation of various variable selection methods. The present section outlines the briefly description of each of the following chapters:

- Chapter 2, provides a detailed examination and presentation of Mixtures of g priors, along with the implementation of a toy example that explores various priors. The objective in this is to compare variable selection methods.

- In Chapter 3, we explain in detail the Fast Bayesian Variable Screening (FBVS) for both simple and multiple regression. We also present simulations and apply this method to real data and we interpret the results.
- Finally , in Chapter 4 , the concluding and discussion section, we present our conclusions and consider opportunities for future research.

Chapter 2

Mixtures of g priors

2.1 Introduction

"Mixtures of g-priors" refers to a type of prior distribution used in Bayesian regression analysis. It is a flexible prior that can accommodate a variety of shapes and can be used to reflect different prior beliefs about the regression coefficients. The "g" in "g-priors" refers to the distribution's shape parameter, which controls the amount of shrinkage or regularization applied to the regression coefficients. Mixtures of g-priors can be useful when there is uncertainty about the appropriate level of shrinkage to apply, as they allow for a blend of different levels of shrinkage to be applied simultaneously.

In linear models, the main objective is to find the most cost-effective way to explain the data using the explanatory variables. This is where Bayesian statistics comes in, as it provides a framework for incorporating prior beliefs about the model and its parameters into the analysis. In this chapter, we will be focusing on a particular type of prior distribution, known as the mixture of g priors, and exploring its applications in linear models and beyond.

2.2 G-prior

The g-prior is a specific form of prior distribution that is often employed to regularize or shrink the parameter estimates towards a specified value (usually zero) in linear regression models. In the context of linear regression, the g-prior for the regression coefficients β is defined as:

$$\beta \sim Normal(0, g(X^T X)^{-1}) \tag{2.1}$$

Here, $Normal(0, g(X^T X)^{-1})$ represents a multivariate normal distribution with mean 0 and covariance matrix $g(X^T X)^{-1}$, where X is the design matrix of predictor variables. The parameter g in the g-prior controls the strength of the shrinkage: larger values of g result in stronger shrinkage towards zero.

Choosing an appropriate value for g is crucial, and it often involves prior knowledge about the scale of the regression coefficients. If you have little prior knowledge, you might choose a small value for g, which leads to weak shrinkage. On the other hand, if you believe that most coefficients should be close to zero, you might choose a larger value for g to enforce stronger shrinkage.

The g-prior is just one example of how prior information can be incorporated into Bayesian analysis, and its specific form is often chosen based on the problem at hand and the researcher's beliefs about the parameters. The choice of prior can significantly influence the results of Bayesian analysis, so it's important to carefully consider and justify your choice of prior distribution.

2.3 Zellner's g priors

In Bayesian linear regression, a common approach is to specify a prior distribution on the regression coefficients that incorporates prior beliefs or information about the data-generating process. Zellner's g prior(30) is a popular choice for this purpose, especially in high-dimensional settings where there are many potential predictors. Due to its ability to greatly simplify posterior computations and decrease the number of prior variance parameters that need to be provided to just one, this prior has been used extensively.

For Gaussian regression, Zellner proposed a prior of the Normal–Gamma family:

$$p(\varphi) = \frac{1}{\varphi} \tag{2.2}$$

$$\beta|\varphi \sim Normal(\beta_a, \frac{g}{\varphi}\left(X^T X)^{-1}\right) \tag{2.3}$$

In equation (2.2) , φ indicates the parameter that represents the precision parameter of a prior distribution. The notation $p(\varphi)$ represents the probability density function (PDF) of φ, which describes our prior beliefs or knowledge about the value of φ before observing any data. The specific form of this prior is the reciprocal of φ, which suggests that smaller values of φ are more probable.

As for the second equation (2.3) , β is the vector of regression coefficients, X is the design matrix of the regression model, X^T is the matrix product of the transpose of X and X. The hyperparameter g controls the strength of the prior. The term "strength" can be interpreted as the degree of regularization imposed by the prior distribution, where a stronger (i.e., more concentrated) prior leads to stronger regularization and a weaker (i.e., wider) prior leads to weaker regularization. More analytically, g is a scalar hyperparameter that controls the overall scale of the prior covariance matrix of the regression coefficients.

In hypothesis testing, when comparing two hypotheses, $H_0 : \beta = \beta_0$ and $H_1 : \beta \in \mathbb{R}^k$, Zellner proposed a prior distribution for the parameter β under H_1. This prior distribution is characterized by a prior mean, denoted as β_a, and a prior covariance matrix that is a scalar multiple (g) of the Fisher information matrix. The value of g depends on the observed data through the design matrix X.

Zellner suggested setting $\beta_a = \beta_0$ in the prior distribution for β under H_1. This means that the anticipated value of β, based on imaginary data, is assumed to be equal to the specified value β_0. By doing this, Zellner derived expressions for the Bayes factor to quantify the evidence in favor of H_1 over H_0 based on this prior distribution.

For testing precise hypotheses, Zellner (1986)(30) derived Bayes factors using g priors; however, he did not expressly take into account nested models, where the null hypothesis limits the values for a subvector of β. By using a flat prior on the regression coefficients that are shared by both models and Zellner's g prior for the regression parameters that are present only in the more complicated model, we can adapt Zellner's g prior for testing nested hypotheses.

The Bayes factor, which is a measure of the relative strength of evidence supporting one hypothesis over another, allows for the expression of posterior probabilities of models.

$$p(M_\gamma|Y) = \frac{p(M_\gamma)BF[M_\gamma : M_b]}{\sum_{\gamma'} p(M_{\gamma'})BF[M_{\gamma'} : M_b]}$$

Where,

$$BF[M_\gamma : M_b] = \frac{p(Y|M_\gamma)}{p(Y|M_b)}$$

And as a result we can generalize this to any comparison of models from the calculation of Bayes factor:

$$BF\left(M_\gamma : M_{\gamma'}\right) = \frac{BF\left(M_\gamma : M_b\right)}{BF\left(M_{\gamma'} : M_b\right)} \tag{2.4}$$

As long as the priors for each model's parameters are defined separately and are independent of the comparison being made, the choice of the base model M_B is theoretically entirely arbitrary. The only options for M_B are the null model and the full model, which causes each combination of M_γ and M_B to be a pair of nested models. The decision to use M_N (the null model) as the base model will be referred to as the null-based strategy. The full model (M_F) is used as the base model in the full-based strategy.

Null-Based Bayes Factors

Zellner's g prior has been shown to have some advantages in Bayesian hypothesis testing, particularly in the calculation of Bayes factors and the estimation of marginal likelihood of models. The g prior allows for the incorporation of prior information about the magnitude of the regression coefficients, which can help to reduce the impact of outliers and improve the accuracy of model selection. Additionally, the g prior can be used to define a family of priors that can adapt to the complexity of the model, making it a versatile tool in Bayesian inference. Overall, the use of Zellner's g prior can lead to more robust and accurate inference in a variety of statistical applications.

$$p(Y|M_\gamma, g) = \left(\frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{n}(\sqrt{\pi})^{n-1}} \right) ||Y - \bar{Y}||^{1-n} \frac{(1+g)^{(n-1-p_\gamma)/2}}{[1+g(1-R_\gamma^2)]^{(n-1)/2}}$$

where R_γ^2 is the ordinary coefficient of determination of regression model M_γ. Now are following the null and the full based strategy.

When computing Bayes factors and model probabilities using the null-based method, we compare each model M_γ with the null model M_N using the hypotheses $H_0 : \beta_\gamma = 0$ and $H_1 : \beta_\gamma \in R$. M_γ and M_N share the intercept as a common parameter which is centered without losing generality. We can assume independence of the null model M_N from the g-prior because when we impose the null hypothesis $H_0 : \beta_\gamma = 0$, the corresponding restricted model has $p_\gamma = 0$ parameters. This implies that the R_γ matrix is a 0×0 matrix, which in turn means that $R_\gamma^2 = 0$. As a result, the value of g does not affect the posterior distribution of the null model, making it independent of the g-prior. Based on this assumption of independence, we can compute the marginal likelihood of both the null model and the model M_γ.

$$p(\alpha, \varphi | M_\gamma) = \frac{1}{\varphi} \tag{2.5}$$

$$\beta_\gamma | \varphi, M_\gamma \sim Normal(0, \frac{g}{\varphi}(X_\gamma^T X_\gamma)^{-1}) \tag{2.6}$$

Are the default priors for α, β_γ and φ under M_γ. Using the Bayes factor, we can compare the strength of evidence for M_γ versus the null model. The Bayes factor is given by the ratio of the marginal likelihoods of the two models, which can be calculated using the g-prior. Therefore, the Bayes factor of M_γ versus M_N is determined by the marginal likelihoods of the two models under the g-prior, and the independence assumption of the null model from the g-prior simplifies the computation of the Bayes factor.

$$BF_{M_\gamma, M_N} = \frac{f(Y|M_\gamma, g)}{f(Y|M_N, g)} = \frac{\frac{(1+g)^{(n-1-p_\gamma)/2}}{[1+g(1-R_\gamma^2)]^{(n-1)/2}}}{\frac{(1+g)^{(n-1)/2}}{(1+g)^{(n-1)/2}}}$$

$$BF_{M_\gamma, M_N} = (1+g)^{(n-1-p_\gamma)/2}\left[1 + g(1 - R_\gamma^2)\right]^{(1-n)/2} \tag{2.7}$$

Zellner's g prior can also be used for hypothesis testing in the context of linear regression models. In this approach, a null model is defined as a restricted version of the full model, where some of the regression coefficients are set to zero.

Under the null model, the Zellner's g prior can be used to obtain a Bayes factor, which is a measure of the evidence in favor of the null hypothesis relative to the alternative hypothesis. The Bayes factor is defined as the ratio of the marginal likelihoods of the null model and the full model, integrated over the model parameters.

If the Bayes factor is greater than one, it indicates evidence in favor of the null hypothesis, while a value less than one indicates evidence in favor of the alternative hypothesis. The strength of evidence is typically interpreted using Jeffreys' scale, where values between 1 and 3 provide "weak" evidence, values between 3 and 10 provide "substantial" evidence, and values greater than 10 provide "strong" evidence.

Full-Based Bayes Factors

To compare model M_γ with covariates X_γ to the full model, we will divide the design matrix for the full model into $X = [1, X_\gamma, X_{-\gamma}]$. Thus, the full model M_F, expressed in partitioned form, is given by:

$$Y = 1\alpha + X_\gamma \beta_\gamma + X_{-\gamma}\beta_{-\gamma} + \epsilon \tag{2.8}$$

Where $X_{-\gamma}$ denotes the columns of design matrix that model M_γ ignores and for $\epsilon \sim Normal(0, \sigma^2)$. Under the null hypothesis, $H_0 : \beta_{-\gamma} = 0$ we obtain the M_γ model, while the alternative hypothesis $H_1 : \beta_{-\gamma} \in \mathbb{R}^{p-p_\gamma}$ the model we assume is the full, with α and β common parameters. Zellner and Siow(31) in 1980 assumed a block-orthogonal parameterization such that $1^T[X_\gamma, X_{-\gamma}] = 0$ and $X_\gamma^T X_{-\gamma} = 0$, to support using α and β_γ as shared parameters across both models. With the above theory they conclude to the following g priors for the Full-based strategy.

$$M_\gamma : p(\alpha, \varphi, \beta_\gamma) \propto \frac{1}{\varphi} \tag{2.9}$$

$$M_F : p(\alpha, \varphi, \beta_\gamma) \propto \frac{1}{\varphi}, \beta_{-\gamma}|\varphi \sim Normal(0, \frac{g}{\varphi}(X_{-\gamma}^T X_{-\gamma})^{-1}) \tag{2.10}$$

As a result the Bayes factor corresponds to:

$$BF_{M_\gamma, M_F} = (1+g)^{(p-n+1)/2} \left[1 + g \frac{1 - R_F^2}{1 - R_\gamma^2} \right]^{(n-p_\gamma-1)/2}$$

Where R^2_F and R^2_γ are, respectively, the coefficients for determining models M_F and M_γ. Because the prior distribution for in M_F depends on M_γ , which varies with each model comparison, it should be emphasized that, unlike the null-based strategy, the full-based approach does not result in a coherent prior specification for the full model.

The null-based strategy assumes that the model with only the intercept and no covariates (null model) is the appropriate reference model. This strategy involves comparing each model of interest to the null model to calculate the Bayes factor. The null-based strategy is simpler and computationally more efficient since it requires only a single model and a simpler prior. However, the null-based strategy is often criticized for being too simplistic and not taking into account the complexity of the models being compared.

The full-based strategy, on the other hand, involves comparing each model of interest to a full model that includes all the covariates. This strategy can be more accurate and informative, especially when the models being compared are complex and contain many covariates. However, the full-based strategy requires more computational resources and may not be feasible in some situations where the number of covariates is large. Overall, the choice between the null-based and full-based strategy depends on the specific research question and available resources.

Lindley's - Bartlett's Paradox

Lindley's - Bartlett's Paradox refers to the apparent contradiction between the finding that individual decision-makers often make more accurate predictions when using averaged forecasts, compared to using their own unaveraged forecasts. In other words , it examines the phenomenon where combining information from multiple sources can lead to a less accurate decision than relying on information from a single source.

The paradox was named after Sir Francis Bartlett(4), a British psychologist and statistician, and Dudley Lindley(20), a British statistician and decision theorist. Since its discovery in the 1950s, it has been extensively studied in the areas of statistics, decision theory, and psychology.

To be more specific , regarding mixtures of g priors and variable selection , even if g prior is deliberately set to be very large and non-informative, the posterior can be reasonable for inference under a particular model. However, in terms of model selection, this is typically a mistake.

The Lindley-Bartlett paradox may favor the smaller or null model over a larger model, even when the data provides evidence for the larger model. This occurs in the special case where the sample size (n) and the number of parameters (p_γ) are constant, and the hyper-parameter g tends to infinity. As g increases, the Bayes factor for comparing the larger model(M_γ) and the null model (M_N) approaches zero, which forces the Bayes factor to support the null model as the most likely despite the information provided by the data.

In particular, when the prior has a large spread induced by the non-informative choice of g it can force the Bayes factor to favor the null model, even if the data provides evidence for a more complex model. This is because the Bayes factor measures the relative evidence for different models, and the large spread of the non-informative prior can dominate the information in the data, leading to a preference for the simplest model. This highlights the importance of careful consideration of the prior choice in Bayesian statistics, as the choice of prior can have a significant impact on the results and conclusions.

Information Paradox

The paradoxes of g priors, also include the information paradox. The information paradox arises when a model M_γ fits a set of data perfectly, such that the coefficient of determination R_γ square tends to 1

$$\lim_{n\to\infty} {R_\gamma}^2 = 1$$

and the classical statistical testing F_γ tends to infinity

$$\lim_{n\to\infty} F_\gamma = \infty$$

when the sample size n and the number of predictors p_γ are held constant.

In this case, we may expect a greater posterior probability for the model M_γ, resulting in a Bayes factor that compares the model M_γ to the null model M_N to tend to infinity as the evidence against the null model accumulates.

Nevertheless, when we use a fixed choice of g, the Bayes factor discussed in the above section converges to a constant value $(1+g)^{(n-p_\gamma-1)/2}$ as the coefficient of determination ${R_\gamma}^2$ tends to 1, which is contrary to our expectations. This convergence leads us contrary to the expected expectations and raises questions about the use of g priors in Bayesian hypothesis testing, as it suggests that the prior distribution may have a stronger influence on the results than the data itself.

Under the assumption of uniform prior model probabilities, the choice of the hyperparameter g in Bayesian model selection methods has a significant impact on model selection. The hyperparameter g controls the weight given to the complexity penalty term in the marginal likelihood calculation, and therefore has a direct impact on the resulting Bayes factors and posterior model probabilities. Specifically, larger values of g will lead to stronger complexity penalties and a greater tendency to select simpler models with fewer but large coefficients. On the other hand, smaller values of g will place less emphasis on the complexity penalty term and may result in more complex models being selected with small coefficients. Therefore, the choice of g should be carefully considered in order to balance the trade-off between model complexity and model fit (George and Foster, 2000)(12). Recommendations for g have included the following:

Unit Information Prior

From the perspective of classical statistics, certain goodness-of-fit criteria are used which lead to model selection. One such criterion is the coefficient of determination, and another interesting criterion is the Bayesian Information Criterion (BIC), which takes into account that the coefficient of determination increases monotonically as the number of parameters increases. Kass and Wasserman (1995)(17) proposed the use of prior distributions that provide information about the parameter equal to the information contained in one observation.

Thus, in the case of normal linear regression, they suggested the use of the unit information prior corresponds to taking g = n. With this choice, we are led to results similar to those of the BIC criterion.

AIC and BIC

AIC and BIC re commonly employed to compare different models and determine which one is the most appropriate for a given dataset.AIC is a measure of the goodness of fit of a statistical model. It balances the trade-off between the goodness of fit (how well the model explains the data) and the complexity of the model (how many parameters are used). AIC is calculated using the following formula: AIC = -2 ln$(\hat{L}) + 2p$, where $\hat{L}$: Maximum likelihood estimate of the model and p: Number of parameters in the model.

BIC, also known as Schwarz Information Criterion (SIC) or sometimes as Schwarz's Bayesian Criterion (SBC), is another criterion for model selection. Like AIC, BIC penalizes complex models to avoid overfitting. BIC is calculated using the following formula: $BIC = -2\ln(\hat{L}) + p\ln(n)$, where n is the sample size.

Differences between AIC and BIC:

- **Penalty for Model Complexity:**
  - **AIC:** Penalizes complex models with a penalty term of $2p$.
  - **BIC:** Heavily penalizes complex models with a penalty term of $p\ln(n)$.
- **Sample Size Influence:**
  - **AIC:** Does not directly account for sample size in the penalty term,and, in turn, select more complex models.
  - **BIC:** Incorporates sample size with a penalty term of $p\ln(n)$.

In summary, AIC and BIC are both useful tools for model selection. AIC is more lenient towards complexity, which might be beneficial when dealing with larger datasets . While BIC provides a stronger penalty for complexity, making it useful, especially for smaller datasets where overfitting is a significant concern. The choice between AIC and BIC often depends on the specific dataset and the balance between model fit and complexity that is desired for the given problem.

Risk inflation criterion

Foster and George (1994)(11) studied prior distributions for model selection based on the risk inflation criterion (RIC) and proposed for the hyperparameter g the value $g = p^2$ from a minimax perspective.

Benchmark prior

Fernandez et al. (2001)(10) conducted a study with various choices for the hyperparameter g that depended on the sample size n and the number of parameters, ultimately selecting g = max(n,p^2) which is referred to as the "Benchmark Prior" or "BRIC" as it combines the BIC and RIC criteria.

Local empirical Bayes

The local empirical Bayesian approach is based on estimating the hyperparameter separately for each model. Using the marginal likelihood and integrating over all the parameters, an empirical Bayesian estimate of the hyperparameter is given by estimating the maximum value of the marginal likelihood subject to the constraint that it is non-negative. This results in the following estimate for g

$$\hat{g}_{\gamma}^{\text{EBL}} = \max\{F_{\gamma} - 1, 0\} \tag{2.11}$$

where

$$F_{\gamma} = \frac{R_{\gamma}^2 / p_{\gamma}}{(1 - R_{\gamma}^2)/(n - 1 - p_{\gamma})}$$

is the usual F statistic for testing β_γ=0. It is easy to estimate an asymptotic SE for g from the observed data.

Global empirical Bayes

According to this choice, the estimation for the hyperparameter is common for all models and arises from the Bayesian model averaging using the posterior model probabilities.

$$\hat{g}^{\text{EBG}} = \arg\max_{g>0} \sum_{\gamma} p(M_\gamma) \frac{(1+g)^{(n-1-p_\gamma)/2}}{[1+g(1-R_\gamma^2)]^{(n-1)/2}}$$

The above estimate is not computable in closed form, however, numerical methods have been developed that could be used (George and Foster 2000)(12).

The Unit Information Prior, Risk Inflation Criterion, and Prior Benchmark prior distributions do not solve the problem of the Information Paradox, as the choice of hyperparameter does not depend on the data information. However, under the two Bayesian empirical methods we just mentioned, we have the following desired behavior in theorem form. The Bayes factor for comparing M_F to M_N goes to ∞ under either the local or the global EB estimate of g in the information paradox context.

2.4 Mixtures of g Priors

The idea behind the g-prior mixtures is based on using the Zellner g-prior by setting a prior distribution on the hyperparameter g. Assuming that the hyperparameter g > 0 is a random variable and setting a prior distribution on it, denoted by $\pi(g)$, which may depend on n, the marginal likelihood of the model $f(Y|M_\gamma)$ is proportional to the Bayes factor in the null-based strategy.

$$BF_{M_\gamma,M_N} = \int_0^\infty (1+g)^{(n-1-p_\gamma)/2} \left[1+(1-R_\gamma^2)g\right]^{(1-n)/2} \pi(g)dg \tag{2.12}$$

The posterior mean of the vector μ_γ under the model selection $M_\gamma \neq M_N$ is given by:

$$\mathbb{E}[\mu|M_\gamma, Y] = 1_n\hat{a} + \mathbb{E}\left[\frac{g}{1+g}|M_\gamma, Y\right] X\gamma\hat{\beta}\gamma \tag{2.13}$$

where $\hat{a}$ is the ordinary least squares estimate of the intercept parameter, $\hat{\beta}_\gamma$ is the maximum a posteriori estimate of the regression coefficients β_γ, X_γ is the design matrix for model M_γ, and $\mathbb{E}\left[\frac{g}{1+g}|, M_\gamma, Y\right]$ is the posterior expectation of the hyperparameter g under model M_γ. The posterior mean for β_γ under a selected model with fixed g prior is a linear shrinkage estimator with a fixed shrinkage factor g/(1 + g). The optimal Bayes estimate of μ under squared error loss is the posterior mean under model averaging given by:

$$\mathbb{E}[\mu|Y] = 1_n\hat{a} + \sum_{\gamma:M_\gamma \neq M_N} p(M_\gamma|Y)\mathbb{E}\left[\frac{g}{1+g}|M_\gamma, Y\right] X\gamma\hat{\beta}\gamma \tag{2.14}$$

Given that the hyperparameter g appears not only in the Bayes factors and model probabilities as well as in the posterior means and predictions, the choice of the prior distribution for the parameter g is crucial to avoid difficult calculations. Below, we will see two options for choosing the prior distribution on the hyperparameter g. The first one is Zellner–Siow's Cauchy prior (Zellner and Siow 1980)(31), which is achieved by applying an Inverse-Gamma prior to g and the second modification is and the hyper-g prior an extension of Strawderman's (1971)(28) prior to regression environment.

Zellner–Siow Priors

In the hypothesis testing for univariate means Jeffrey's largely dismissed normal priors due to the fact that they led to paradoxes of Bayes factor paradoxes. He also proposed the use of Cauchy prior which is satisfied the basic consistency requirements for hypothesis testing. Specifically, Zellner-Siow (1980)(31) proposed multivariate Cauchy prior distributions for the regression coefficients which can be solution to the problem of the mean. When two nested models are being compared the common coefficients are given a flat prior, and the remaining parameters are given a Cauchy prior. In the null-based approach, the prior on

(α,φ) is given by Liang and a multivariate Cauchy centered at the null model, β_γ=0.

$$\pi(\beta_\gamma|\varphi) \propto \left(\frac{\Gamma(p_\gamma/2)}{\pi^{p_\gamma/2}}\right) \frac{X_\gamma^T X_\gamma}{n/\varphi}^{1/2} \left(1 + \frac{\beta_\gamma^T X_\gamma^T X_\gamma \beta_\gamma}{n/\varphi}\right)^{-p_\gamma/2} \tag{2.15}$$

The Zellner-Siow prior distribution, however, did not become as well-known as the g prior distribution in Bayesian variable selection, and the explanation lies in the fact that it presents several difficulties in calculating the marginal likelihood since there are no available closed forms of computation. Although several approximation attempts were made, such as Laplace approximations, it was observed that as the dimension of the model increases, the accuracy of these approximations decreases.

Given that the Cauchy distribution can be represented as a mixture of normal's, a mixture of g prior distributions with inverse gamma prior distributions on g specifically, $\text{Inverse} - \text{Gamma}\left(\frac{1}{2}, \frac{n}{2}\right)$ priors can be used. In other words, we can take:

$$\pi(\beta_\gamma|\varphi) \propto \int Normal(\beta_\gamma|0, \frac{g}{\varphi}(X_\gamma^T X_\gamma)^{-1})\pi(g)\, dg \tag{2.16}$$

with

$$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/2g} \tag{2.17}$$

With the aforementioned prior distribution, after integrating with respect to θ_γ when computing the marginal likelihood of the model, due to the proportionality described, we can easily see that by substituting the prior distribution, only a one-dimensional integral with respect to g remains, which is independent of the dimension of the model. This integral can be solved using standard integration techniques or Laplace's approximation. Laplace's method leads to easy computations for approximating the marginal likelihood $f(Y|M_\gamma)$ as well as the posterior expected value of the factor g/(1+g) which is necessary for predictions, since the posterior marginal value for the hyperparameter g is a solution of a cubic equation (Feng Liang et.al, 2007)(19).

Hyper g priors

The hyper-prior distribution, introduced by Liang et al. (2008)(19), encompasses a family of priors that have been extensively studied in various statistical problems. These priors, including those used by Strawderman (1971) (28) and investigated by Cui and George (2007)(7). There is an observable enhancement in the mean square risk concerning the ordinary maximum likelihood estimates within the context of the normal means problem. They are particularly relevant in cases involving variable selection when the error variance is known. By treating the hyperparameter g as a random variable, the hyper-prior distribution allows for a Bayesian approximation of the model's marginal likelihood in a closed form, enabling convenient and analytical computation. Overall, this approach extends the classical g prior distribution and proves valuable for linear models and related analyses. As Liang et al. (2008)(19) introduced a family of priors on g which is a proper distribution for $\alpha > 2$:

$$\pi(g) = \frac{(a-2)}{2}(1+g)^{-a/2}, \quad g > 0 \tag{2.18}$$

When the value of a is less than or equal to 2, the prior distribution $\pi(g)$ with a proportional relationship to $(1+g)^{-\alpha/2}$ is considered to be improper. In this case, both the reference prior and the Jeffrey's prior correspond to a equaling 2.

When the value of α is greater than one and less than or equal to two ($1 < \alpha \leq 2$), it becomes evident that the marginal density can be derived from the given expression. The quantity represented by (2.18) is categorized as finite. As a result, the resulting posterior distribution is regarded as proper. Despite the fact that the selection of the interval $1 < \alpha \leq 2$ results in the need to include g in the calculation.

In the context of the null model, the concern surrounding arbitrary constants of proportionality results in the occurrence of indeterminate Bayes factors. Due to this particular circumstance, it is imperative that we restrict the focus of attention in equation (2.18) to values greater than 2. This measure aims to refine the scope of the analysis and achieve more precise outcomes.

A deeper understanding of hyperparameter selection can be achieved by examining the prior associated with the shrinkage factor expressed as $g/(1+g)$, where :

$$\frac{g}{1+g} \sim \text{Beta}(1, \frac{\alpha}{2} - 1) \tag{2.19}$$

which is a Beta distribution with mean $\frac{2}{\alpha}$.

Specifically , substituting $a = 4$, we have:

$$\frac{g}{1+g} \sim \text{Beta}\left(1, \frac{4}{2} - 1\right)$$

$$\frac{g}{1+g} \sim \text{Beta}\,(1, 2-1)$$

$$\frac{g}{1+g} \sim \text{Beta}\,(1, 1)$$

The mean of a Beta distribution with parameters α and β is given by $\frac{\alpha}{\alpha+\beta}$. In this case, $\alpha = 1$ and $\beta = 1$, so:

$$\text{Mean of Beta}(1,1) = \frac{1}{1+1} = \frac{1}{2}$$

Therefore, when $a = 4$, $\frac{g}{1+g}$ follows a Beta(1, 1) distribution which is a uniform distribution.

Table 2.1: Prior on Shrinkage Factor for Different Values of a

	Prior Description	Prior Probability
$\alpha = 3$	Most mass near 1	$P(\text{shrinkage factor} > 0.80) = 0.45$
$\alpha = 4$	Uniform	–
$\alpha > 4$	More mass near 0	–

Therefore, values $2 < \alpha \leq 4$ are considered reasonable to work with and provide the main advantage of the hyper-g prior distribution, leading to a closed-form posterior distribution for the hyperparameter g given the model M_γ.

$$p(g|Y,M_\gamma) = \frac{p_\gamma + \alpha - 2}{2{}_2F_1\left((n-1)/2, 1; (p_\gamma+\alpha)/2; {R_\gamma}^2\right)} (1+g)^{(n-1-p_\gamma-\alpha)/2}[1+g(1-{R_\gamma}^2)]^{(1-n)/2} \tag{2.20}$$

Where ${}_2F_1(a,b;c;z)$ in the normalizing constant is the Gaussian hypergeometric function (Abramowitz and Stegun 1970)(1). The hypergeometric function ${}_2F_1$(a, b; c; z) is a special mathematical function. It is represented as a power series and is a solution to the hypergeometric differential equation. The parameters a, b, c, and z define the behavior of this function.

The integral representing ${}_2F_1(a,b;c;z)$ is tractable, under these circumstances: for real $|z| < 1$ with $c > b > 0$, and for $z = \pm 1$ only if $c > a + b$ and $b > 0$. As the normalizing constant in the prior on g is also a special case of the ${}_2F_1$ function with z = 0, we refer to this as the hyper-g priors. Another interesting appearance of the function ${}_2F_1(a,b;c;z)$ is in the Bayes factor that we use to compare the model M_γ with the null model. Specifically, the normalization constant in the posterior distribution for the hyperparameter g leads to the following:

$$\begin{aligned} BF[M_\gamma : M_N] &= \frac{\alpha-2}{2}\int_0^\infty (1+g)^{(n-1-p_\gamma-\alpha)/2}[1+g(1-R_\gamma^2)]^{(1-n)/2} dg \\ &= \frac{\alpha-2}{p_\gamma+\alpha-2}\,{}_2F_1\left(\frac{n-1}{2}, 1; \frac{p_\gamma+\alpha}{2}; R_\gamma^2\right) \end{aligned} \tag{2.21}$$

Which is tractable and the posterior mean of g for a>3 is finite under M_γ is the following:

$$\mathbb{E}[g|M_\gamma, Y] = \frac{2}{p\gamma + \alpha - 4} \frac{{}_2F_1((n-1)/2, 2; (p_\gamma+\alpha)/2; R_\gamma^2)}{{}_2F_1((n-1)/2, 1; (p_\gamma+\alpha)/2; R_\gamma^2)} \tag{2.22}$$

The ${}_2F_1$ function can also be used to represent the expected value of the shrinkage factor under each model. Unlike the regular g prior, the ${}_2F_1$ function results in nonlinear data-dependent shrinkage.

$$\mathbb{E}\left[\frac{g}{1+g} \mid Y, M_\gamma\right] = \frac{\int g(1+g)^{(n-1-p_\gamma-\alpha)/2-1}\left[1+(1-R_\gamma^2)/g\right]^{(1-n)/2} dg}{\int (1+g)^{(n-1-p_\gamma-\alpha)/2}\left[1+(1-R_\gamma^2)g\right]^{(1-n)/2} dg} = \frac{2}{p_\gamma+\alpha}\frac{{}_2F_1((n-1)/2,2;(p_\gamma+\alpha)/2+1;R_\gamma^2)}{{}_2F_1((n-1)/2,1;(p_\gamma+\alpha)/2;R_\gamma^2)} \tag{2.23}$$

For the computation of the Gaussian hypergeometric function, various subroutines have been developed. However, they often encounter computational difficulties for large n and ${R_\gamma}^2$. Improved numerical methods include Laplace approximations (Tierney and Kadane 1986)(29).

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a}dt$$

2.4.1 Toy-example

The ‘Ozone’ data from mlbench package contains the Los Angeles ozone pollution data collected in 1976. It is now up to us to find out which of the available variables would suit best to predict the ozone reading with maximum accuracy. The objective of this analysis is to accurately predict the ‘daily maximum one-hour average ozone reading’, using linear regression models.

The BAS package is used in linear models for Bayesian model averaging with stochastic or deterministic sampling without replacement from the posterior distribution of the models. The form of the prior distributions used for the coefficients is based on Zellner’s g-prior and g-prior mixtures. For fewer than 20-25 features, it examines all models as is the case in the current example; otherwise, it performs random or deterministic sampling without replacement in the model space.

The posterior means results for the coefficients are presented comparatively for all these cases in the table. The third decimal place has been rounded off, we observe that the marginal posterior estimates for the means of the corresponding coefficients under the prior

Table 2.2: Marginal posterior means for the coefficients (2^pcandidate models)

	AIC	*BIC*	*g*	*JZS*	*Zs-null*	*Zs-full*	*Hyper-g*	*Hyper-g-Laplace*	*Hyper-g/n*	*EBL*	*EBG*
Intercept	11.37	11.37	11.37	11.37	11.37	11.37	11.37	11.37	11.37	11.37	11.37
Pressure Height	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
WindSpeed	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Humidity	0.09	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.11	0.10	0.10
Temperature Sandburg	0.13	0.08	0.07	0.08	0.08	0.10	0.08	0.08	0.08	0.08	0.09
Temperature El Monte	0.48	0.45	0.45	0.45	0.45	0.46	0.45	0.45	0.45	0.45	0.45
Inversion Base Height	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pressure Gradient	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Inversion Temperature	−0.08	−0.02	−0.02	−0.03	−0.03	−0.04	−0.03	−0.03	−0.02	−0.03	−0.03
Visibility	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

distributions are with minimal (almost negligible) differences the same.

We observe that both below the AIC criterion selection and below the selection of the prior ZS-full, there is a tendency to select higher-dimensional models, as well as the tendency (compared to the other priors) to assign a higher posterior probability of inclusion to variables that ultimately are not included in the respective model.

The values in the table represent the posterior probabilities of each variable being included in the model under the different priors. Higher values indicate greater likelihood of inclusion. For example, a variable with a high posterior probability under a specific prior is more likely to be important for predicting the target variable using that particular prior. The table provides insights into how different priors influence variable selection and model complexity. In the table, it is presented the posterior probability margin for the coefficients to be non-zero for each of the different prior distribution choices.

Focusing attention on the results for the first proposed model (higher posterior probability model), it is observed that these are similar to the options BIC, EBL, EBG, hyper-g, hyper-g/n, and ZS-null. Under these options, 9 explanatory variables are incorporated into the model. It is worth noting that the intercept term is always included in the list of higher pos-

Table 2.3: Marginal posterior probability p(B!=0) (2^9candidate models)

	AIC	*BIC*	*g*	*JZS*	*Zs-null*	*Zs-full*	*Hyper-g*	*Hyper-g-Laplace*	*Hyper-g/n*	*EBL*	*EBG*
Intercept	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pressure Height	0.80	0.48	0.46	0.48	0.50	0.62	0.53	0.53	0.48	0.53	0.53
WindSpeed	0.29	0.08	0.08	0.09	0.10	0.14	0.11	0.11	0.09	0.11	0.11
Humidity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Temperature Sandburg	0.85	0.53	0.52	0.54	0.56	0.68	0.58	0.59	0.53	0.59	0.59
Temperature El Monte	1.00	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99
Inversion Base Height	0.93	0.66	0.65	0.67	0.69	0.81	0.71	0.71	0.66	0.71	0.73
Pressure Gradient	0.37	0.19	0.19	0.20	0.20	0.26	0.22	0.22	0.20	0.22	0.22
Inversion Temperature	0.53	0.19	0.19	0.21	0.22	0.31	0.24	0.25	0.21	0.24	0.24
Visibility	0.32	0.08	0.08	0.10	0.10	0.15	0.12	0.12	0.09	0.11	0.11

terior probability models under any choice. For the final selection, a comparison was made between the posterior means and standard errors of the coefficients, along with additional diagnostic tests and log marginal likelihood.

The 'Pressure Height' feature highlights a distinct division among the models. Two groups emerge based on the presence of this feature: one group includes models with 'Pressure Height,' and the other group comprises models without it. Notably, the 'Pressure Height' feature is primarily associated with Bayesian priors such as AIC, ZS Null, ZS Full, Hyper-g, Hyper-g Laplace, and EBL. The comparison of these models inside the group will be using posterior means and standard errors and log marginal likelihood. So, the following figures are from BAS package in the environment of R just like the whole analysis.

The log marginal likelihood is a crucial quantity in Bayesian model selection and Bayesian inference. It represents the probability of observing the given data under the model, marginalized over all possible values of the model parameters. In other words, it quantifies how well the model fits the data while accounting for the complexity of the model. The model with the higher log marginal likelihood (or lower negative log marginal likelihood) is generally favored as it better explains the data while penalizing overly complex models.

Table 2.4: Selection of variables with p(B != 0)>0.5 (Median Probability Model)

	AIC	*BIC*	*g*	*JZS*	*Zs-null*	*Zs-full*	*Hyper-g*	*Hyper-g-Laplace*	*Hyper-g/n*	*EBL*	*EBG*
Intercept	*	*	*	*	*	*	*	*	*	*	*
Pressure Height	*				*	*	*	*		*	
WindSpeed											
Humidity	*	*	*	*	*	*	*	*	*	*	*
Temperature Sandburg	*	*	*	*	*	*	*	*	*	*	*
Temperature El Monte	*	*	*	*	*	*	*	*	*	*	*
Inversion Base Height	*	*	*	*	*	*	*	*	*	*	*
Pressure Gradient											
Inversion Temperature	*										
Visibility											

* Statistically significant explanatory variables for p(B != 0)>0.5

Utilizing the log marginal likelihood, our examination reveals that AIC and BIC exhibit comparable values close to zero. Conversely, the rest of the prior selections yield positive differences, barring the ZS full prior which displays a notable distinction. When gauging the options, EBL coupled with 'Pressure Height,' EBG, and the g prior emerge with comparable likelihood values, albeit without 'Pressure Height.' Given the fewer features at hand, opting for the g prior or EBG presents a superior choice. Specifically, the EBG option emerges as the preferable alternative. To further illustrate this point, the following plots from the BAS package with EBG provide additional insight.

In following figure, it is presented a visualization in the model space for each prior separately. This visualization includes the first fifteen models, with the included and excluded explanatory variables in each model, as well as the logarithm of the posterior odds ratio values for each model. Excluded explanatory variables are depicted in black, while any other color represents the explanatory variables included in the model. Variables belonging to the same color group indicate their inclusion in the same model, along with their corresponding posterior probability.

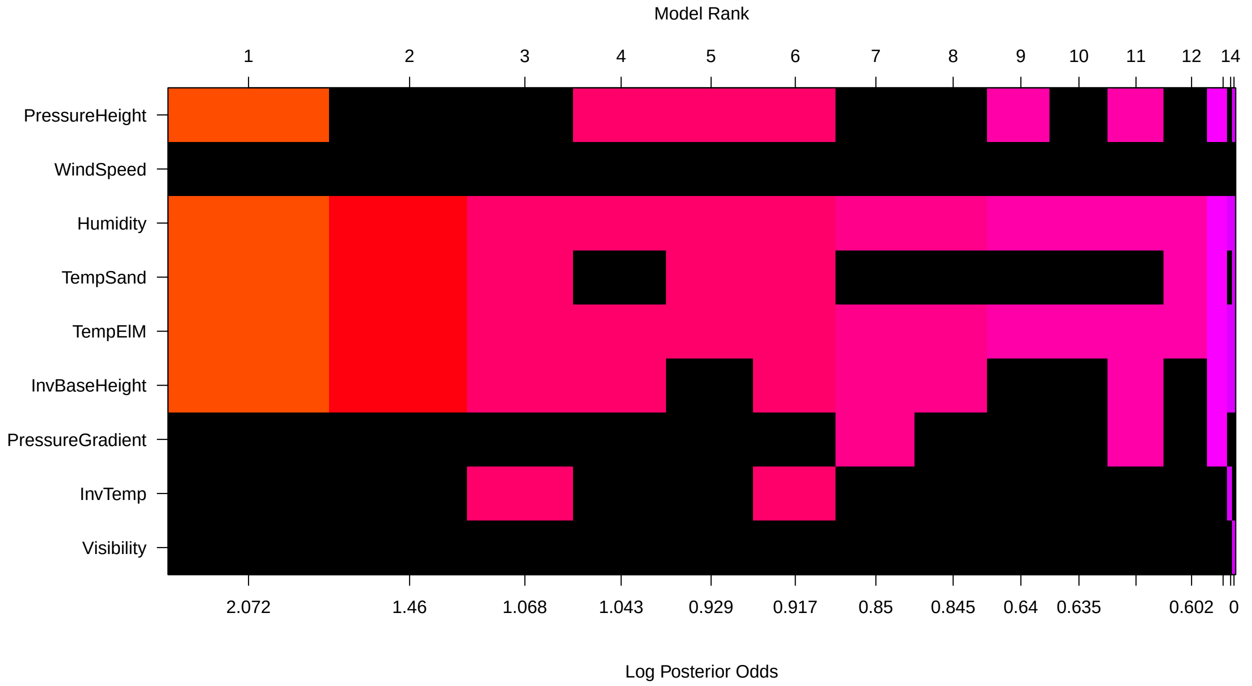

Figure 2.1: Fifteen top model from Bas

2.5 Conclusion

In conclusion, this chapter has encompassed a thorough exploration of Bayesian statistics, dissecting the intricacies of prior distributions and their implications within statistical modeling. The significance of prior distributions in shaping the foundation of Bayesian inference has been expounded upon, highlighting their role as repositories of pre-existing knowledge prior to data observation. The pivotal role of hyperparameters in modulating prior behavior has been illuminated, underscoring the equilibrium between informative and non-informative priors. Concrete instances such as the Zellner-Siow priors and the hyper-g prior have been examined to illustrate the tangible impact of diverse prior choices on posterior inferences. Additionally, the concept of Bayes factors as effective tools for contrasting models has been elucidated, offering a quantitative gauge of relative evidence between competing hypotheses. By embedding these notions within the broader spectrum of Bayesian statistics, a profound grasp of the manner in which prior distributions and Bayes factors steer the course of Bayesian decision-making has been achieved.

Table 2.5: Coefficients for EB-GLOBAL Prior

	post mean	post SD	post p(B $\neq$ 0)
Intercept	11.374	0.318	1.000
PressureHeight	-0.008	0.009	0.535
WindSpeed	0.010	0.066	0.109
Humidity	0.103	0.023	0.999
TemperatureSandburg	0.086	0.086	0.594
TemperatureElMonte	0.450	0.134	0.995
InversionBaseHeight	0.000	0.000	0.726
PressureGradient	0.004	0.011	0.219
InversionTemperature	-0.030	0.093	0.244
Visibility	0.000	0.002	0.112

Looking forward, the subsequent chapter will introduce a groundbreaking algorithm developed by Ntzoufras and Paroli in 2023(26). This algorithm builds upon the principles of the g prior, but with a modification involving the integration of g prior mixtures. This approach will address challenges associated with computational complexity, as exemplified by the hypergeometric function. By unveiling and addressing these challenges. Its expansion into mixtures reflects not only its statistical prowess but also its adaptability in tackling computational obstacles. This chapter highlights the g prior's enduring relevance and its role in shaping the future landscape of Bayesian analysis.

Chapter 3

Fast Bayesian Variable Screening (FBVS)

3.1 Introduction

The method that is being discussed uses mixtures of g-prior specification using both the simple uni-covariate and multiple regression setups and is a quick Bayesian variable selection strategy applied to Normal regression models. The technique makes use of partial and Pearson correlation coefficient thresholds. It is founded on Bayesian ideas that are obtained from posterior model odds and Bayes factor thresholds. The size of the model space can be decreased and inconsequential covariates can be found using this process. As a result, on the resulting condensed model space, more conventional and computationally demanding Bayesian variable selection techniques can be applied without difficulty. Additionally, this method is adaptable and is be expanded to handle different prior settings. This method was initially investigated by Paroli and Ntzoufras, who focused on the selection of the g prior due to its closed-form Bayes factor computations and precise correlation thresholds. It constitutes a highly efficient screening algorithm that relies on only two equations. Notably, this method was studied and developed this year (2023).

3.2 Background

3.2.1 Posterior Inclusion Probability (PIP)

In Bayesian variable selection, the posterior inclusion probability (PIP) is a useful metric for determining how likely it is that a covariate will be included in the model given the observed data. It is written as $P(j = 1|Y)$, where j stands for the covariate X_j's inclusion indicator. The posterior probability of individual models tends to be low when dealing with a vast model space, with the exception of situations where a small number of features exhibit a high signal-to-noise ratio. We may determine the median probability (MP) model $\{X_j : P(\gamma_j = 1|Y) > 0.5 \text{ for } j = 1, \ldots, p\}$ by taking PIPs into account; this model only contains covariates with PIPs greater than 0.5. Under certain conditions the MP model is better than the MAP model in terms of predictive inference (Barbieri & Berger, 2004)(2).

3.2.2 Universal Bayesian Covariate in Simple Regression

Throughout this chapter, there will be a discussion of Fast Bayesian Variable Screening that explicitly utilizes the g prior, along with an exploration of variations involving mixtures of g priors. The first formula will be under g prior like the original form of the algorithm and then will be a generalization under the mixtures of g priors.

Ntzoufras and Paroli had compared two basic regression models using the Bayes factor: the null model M_N and the model M_j, which only incorporates covariate X_j in the linear predictor. The square of the sample coincides with the number of covariates p_γ = 1 and the coefficient of determination R^2_γ for this model. Between Y and X_j, the Pearson correlation coefficient is ρ_j^2. The Bayes factor is the following:

$$B_j^{uni}(g) = BF[M_j : M_N] = \frac{(1+g)^{(n-2)/2}}{[1+g(1-\rho_j^2)]^{(n-1)/2}}, j = 1, \ldots, p \tag{3.1}$$

After Lykou and Ntzoufras (2013)(21) notation 3.1 in their paper, it will be mentioned as the "uni-covariate" Bayes factor. In linear regression setup with only one feature, the posterior probability of M_j can be also treated as the posterior inclusion probability (PIP) of X_j which is the following:

$$PIP_s(j) = \frac{BF[M_j : M_N]}{1 + BF[M_j : M_N]} \tag{3.2}$$

As the Pearson correlation is stronger for each value of n, the Bayes factor offers more evidence against the null model. The uni-covariate Bayes factor is an increasing function of ρ_j for any fixed g.

3.2.3 Correlation Thresholds

This uni-Bayes factor takes a specified value reflecting marginal Bayesian evidence in order to determine a given correlation threshold value. They specifically look for whatever g and ρ_j combination results in a uni-covariate Bayes factor that is less than or equal to a particular Bayes Factor threshold.

This Bayes factor takes a certain value representing marginal Bayesian evidence in order to determine a specified correlation threshold value. In particular, we are looking for which set of values for features, a, and sample size results in a uni-covariate Bayes factor that is below or equal to a certain selection of the Bayes Factor threshold of importance θ.

Therefore, for the model under consideration, in practice, a covariate X_j with a sample squared correlation coefficient ${\rho_j}^2$ such that ${\rho_j}^2 \leq {\rho_\theta}^2$ will not be significant. We call $|\rho_\theta|$ the non-importance correlation threshold as a result.

For all possible values of g, Lykou and Ntzoufras defined the non-important set of correlation coefficients, as the set of correlations that correspond to covariates with a uni-covariate Bayes factor less than θ, some common choices are (1,3,20,150).

$$\rho_\theta : B_j^{uni}(g) \leq \theta, \forall g \geq 0, j = 1, \ldots, p$$

From the equation (2.7) they obtained the upper bound of this set as a function of g and θ, ρ_θ and ρ_θ^2 given by:

$$\rho_\theta^2 = \frac{g+1}{g}\left(1 - [\theta^2(1+g)]^{-1/(n-1)}\right) \tag{3.3}$$

Where $|\rho_\theta|$ is called non-importance correlation threshold, and this was the 1st approach under g-prior. In a special case for the g-prior, choosing g=n (UIP) offers the advantage of yielding results akin to the Bayesian Information Criterion (BIC). This approach, focused on the information contained in a single data point or observation, avoids favoring specific values for the regression coefficients, making it a non-informative prior. By striking a balance between model complexity and regularization, the UIP allows for effective model selection while preserving the statistical integrity of the analysis.

The first approach, as presented in equation (3.3), represents the general and conventional method. The second approach involves utilizing the Unit Information Prior (UIP), while the third approach employs Local Empirical Bayes by maximizing the Bayes Factor (BF). Each approach comes with its own set of advantages and trade-offs.

This expansion builds upon Liang et al.'s equation 17 from their paper(19) on "Mixtures of g Priors for Bayesian Variable Selection." The objective of this study is to expand the algorithmic framework by incorporating the concept of mixtures of g priors. By integrating this concept, we aim to enhance the algorithm's efficacy in selecting relevant variables within a Bayesian context. This research contributes to the advancement of variable selection techniques and their application in diverse fields. Within the scope of this master's thesis, an intricate equation emerges, presenting challenges regarding direct solution. The equation takes the mathematical form:

$$BF_{j0} = \frac{\alpha - 2}{p_\gamma + \alpha - 2}\, {}_2F_1\left(\frac{n-1}{2}, 1; \frac{p_\gamma + \alpha}{2}; R^2_\gamma = \rho\right) \tag{3.4}$$

We assigned the given Bayes Factor as θ and then solved the equation, taking into account the R^2_γ constraint. This process yielded the necessary threshold for our study. The methodology employed for this analysis is elaborated in the subsequent paragraphs. In the context of simple regression, we equate R^2_γ to ρ, representing the correlation coefficient and the correlation threshold for the simple regression.

This equation prominently features a hypergeometric function, denoted as ${}_2F_1$. Hypergeometric functions often lack a straightforward closed-form solution, which complicates the determination of the specific value of R^2_γ. It is essential to acknowledge that obtaining a closed-form solution for R^2_γ remains a challenging endeavor due to the equation's inherent complexity.

Of particular interest is the exploration of the value of R^2_γ and its implications within the equation. However, due to the intricate nature of the hypergeometric function and the mixture of prior distributions in play, directly calculating R^2_γ in a closed form presents difficulties.

In this research, the R package `hypergeo` assumes a critical role within the study's environment. This package offers solutions based on the Hypergeometric and generalized hypergeometric functions as meticulously defined by Abramowitz and Stegun. This capability holds particular significance, as it enables the identification of a crucial threshold. The threshold, in this context, plays a pivotal role in the determination of variables through the extended algorithm, adding a valuable dimension to the expansion of the Bayesian variable selection process.

The subsequent stage of this investigation involves locating the root within a predetermined interval through the utilization of the `uniroot` function, employing various tolerance levels. In essence, the `uniroot` function systematically searches within the specified interval, moving from the lower to the upper bound, to pinpoint a root or zero of the given function concerning its first argument. This step presents an opportunity to establish a threshold utilizing the capabilities of the `hypergeo` package across diverse levels of tolerance. It is essential to acknowledge that the complexity of the hypergeometric function occasionally renders the stability of the interval challenging. In certain instances, the function may struggle to ascertain the root due to the intricate nature of the hypergeometric calculations involved.

The `BMS` package plays a pivotal role in validating the outcomes of our study. This validation process involves employing an alternative library that specializes in computing the Gaussian hypergeometric function for mixtures of g prior proposes. This approach offers an additional layer of confirmation for our derived results. By using this distinct library, we effectively corroborate the root obtained through our initial calculations. This validation step entails assessing the value of the Gaussian hypergeometric function within the specified interval (0, 1). When the root is accurate, the interval associated with this calculated value should approximate zero. In this manner, the utilization of the `BMS` package not only bolsters the reliability of our findings but also verifies the integrity of the root discovered within the Bayesian variable selection process.

In certain scenarios, a direct solution is attainable when we can straightforwardly determine the Hypergeometric Function. However, for specific combinations of features, sample size, and $'a'$, a direct solution cannot be derived using the `uniroot` function to find R^2_γ. This parameter is crucial for both simple and multiple regression. In the case of simple regression, it equates to the coefficient of correlation, while in multiple regression, it differs slightly as it represents partial correlation. Nevertheless, it remains essential in specifying the correlation threshold for the study.

We are exploring combinations of sample size and alpha in the context of simple regression where a solution can be obtained using the `uniroot` function. The figure visualizes the behavior of the hypergeometric function ${}_2F_1$, also known as the Gauss hypergeometric function, as you vary the parameters and the size of the sample. The parameters are chosen from the range [2.1, 4] with increments of 0.1, following the approach by Liang et al. (2008)(19). The sample size, denoted as 'n', ranges from 10 to 100 and the p_γ is equal to 1, resulting in

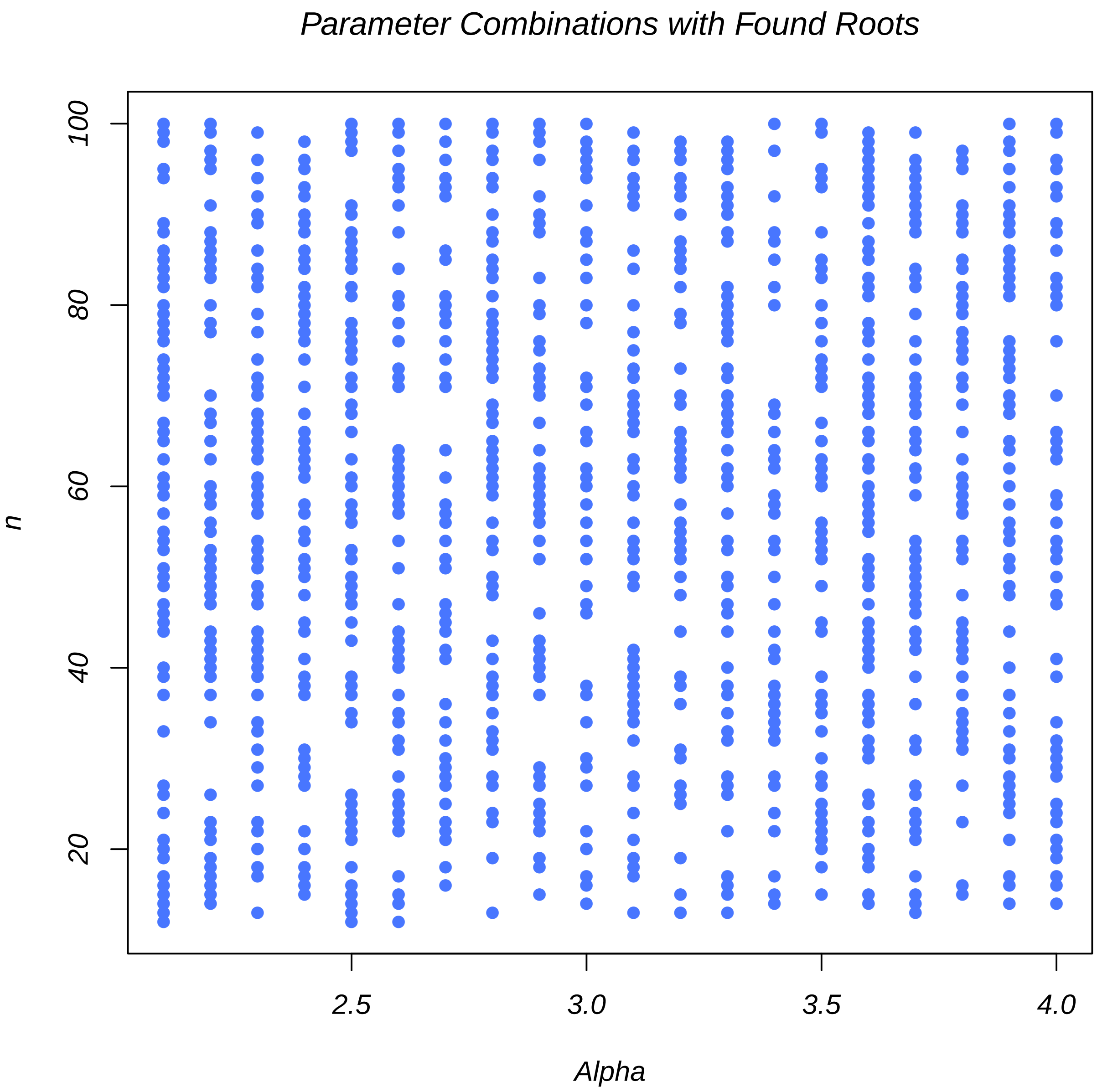

Figure 3.1: Solution of Hypergeometric Function ${}_2F_1$ with varying parameters alpha and n

a total of 1780 combinations for analysis with direct solutions percentage of 54%. However, these instances tend to be concentrated within a particular range of parameter values and smaller sample sizes. Specifically, direct solutions are more prevalent for higher values of 'alpha' and sample sizes.

The percentage of instances with direct solutions rises with increasing sample size n. This observation suggests that as the sample size grows, the likelihood of encountering direct analytical solutions for the hypergeometric function increases. This trend highlights the

Table 3.1: Analysis of Direct solutions by parameter α and Sample Size n

size of sample	Percentage of Direct Solutions
$n = 100$	55%
$n = 1000$	68%
$n = 10000$	77%

growing stability of the function's behavior as sample size expands. This table accentuates the progressive rise in the prevalence of direct solutions with larger sample sizes, which can be crucial information for applications necessitating analytical solutions. This phenomenon suggests that certain parameter ranges, especially when 'alpha' is small, lead to intricate behavior in the hypergeometric function, requiring resorting to numerical or approximated techniques for evaluation.

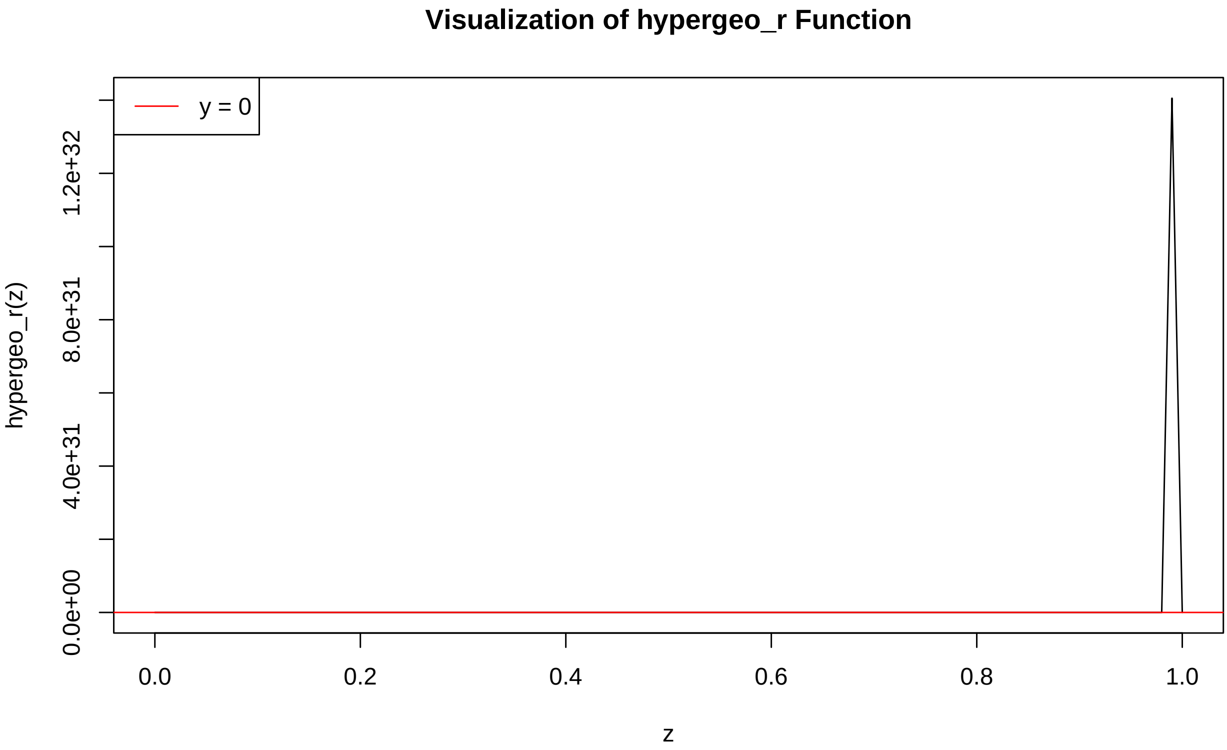

Figure 3.2: Direct solution

In this visual representation, the concept of a direct solution is depicted using the Greek letter λ, symbolizing the solution line. The graph illustrates the attempt to find a solution for a specific problem utilizing the `uniroot` function in R, a common computational approach for finding roots. A direct solution occurs when the solution line (λ) intersects the x-axis at a specific point, indicating a feasible and optimal solution for the given problem. However, the visualization reveals a limitation inherent to the `uniroot` function in R for a specific combination of the parameters. Despite its application, the solution line (λ) fails to intersect the x-axis. This absence of intersection signifies that, with the specific methodology employed (using `uniroot` in R environment), a direct solution cannot be attained. The values for the specific figure $a = 3$, n=50 and $\theta = 1$ and 10 features.

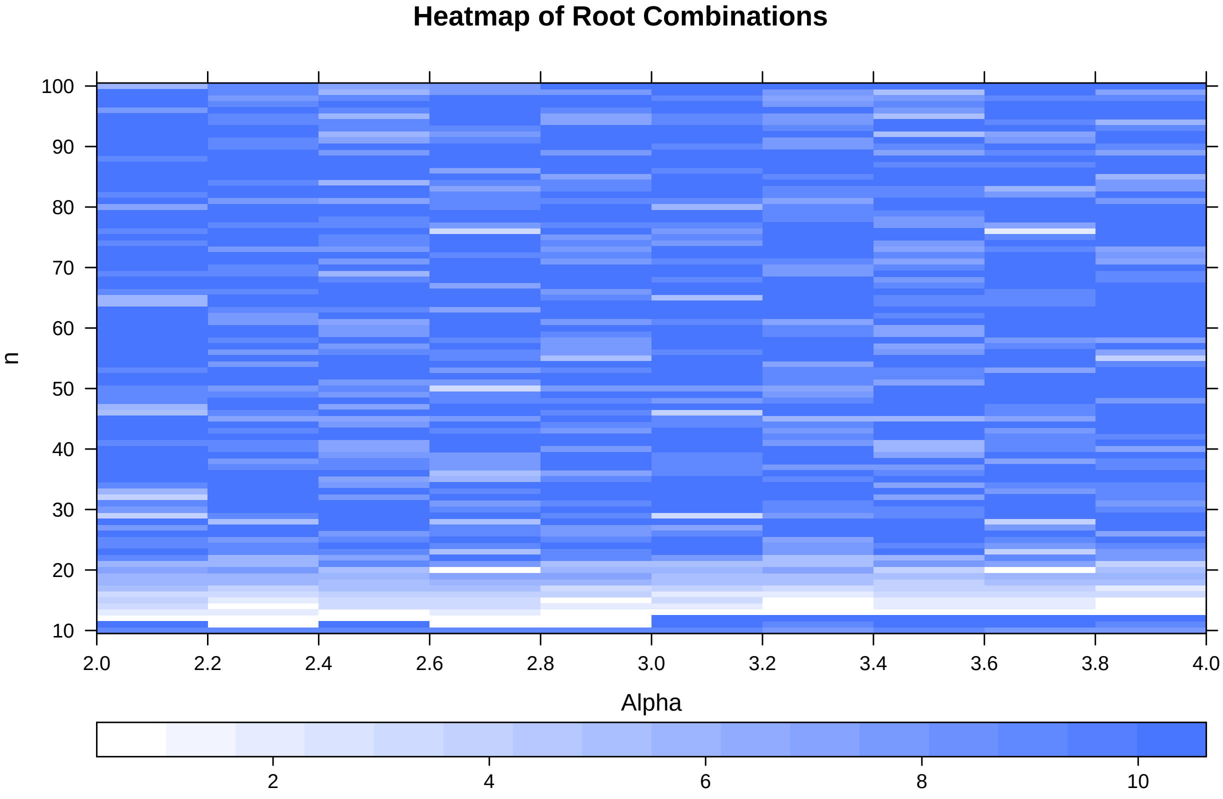

Figure 3.3: Heatmap of Hypergeometric Function ${}_2F_1$ with varying parameters alpha and n

The heatmap visually represents the intricate interplay between sample size (n) and a, capturing the complexity of their relationship within a given dataset. In this heatmap, the amalgamation of colors illustrates various combinations of n and a, with distinctive white spots scattered across the heatmap. Notably, these white spots are particularly prevalent in regions representing very low sample sizes. This intriguing phenomenon occurs because, in many instances, the hypergeometric function lacks solutions when calculated using the `uniroot` function. These white patches serve as a visual testament to the challenging nature of solving the hypergeometric equation for certain combinations of n and a.

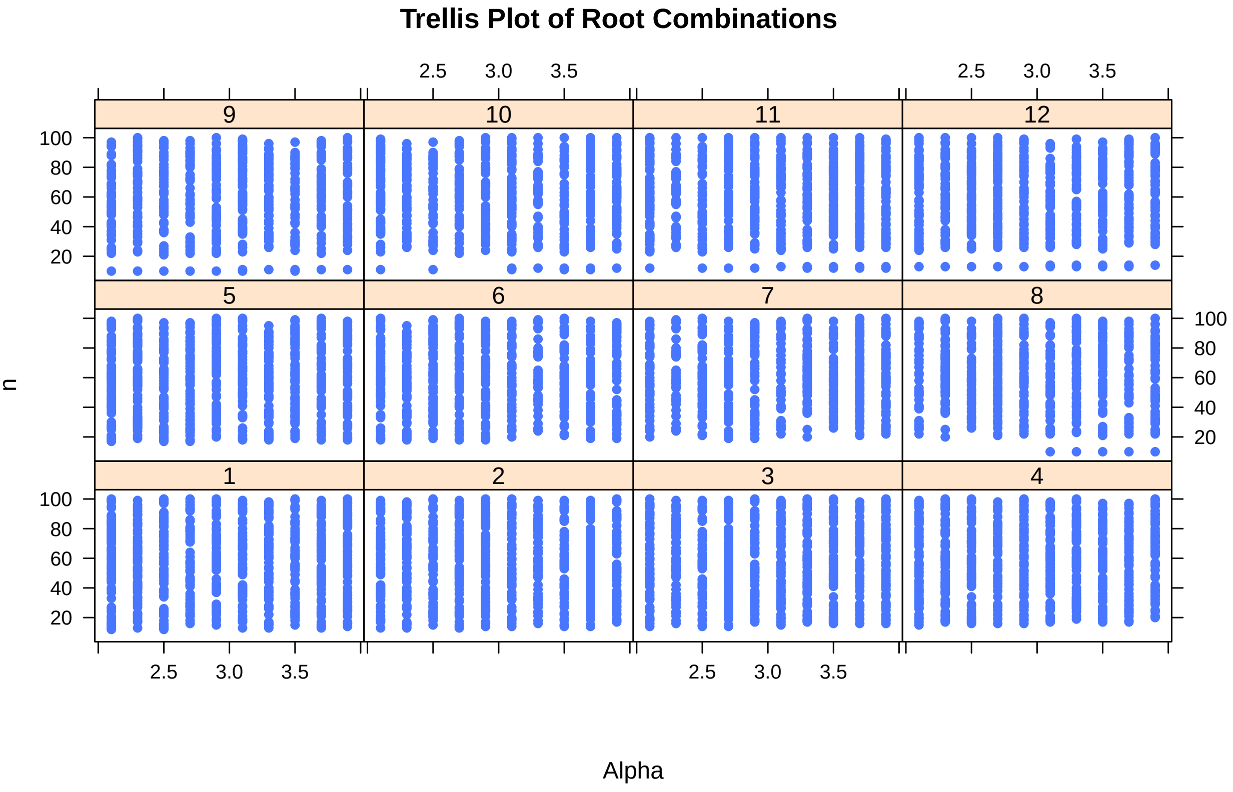

Figure 3.4: Trellis plot of Hypergeometric Function ${}_2F_1$ with varying parameters alpha and n and features

The figure presented following offers a comprehensive representation of the intricate interplay between crucial parameters: sample size ('n'), parameter alpha ('α'), and the number of potential features ('p'). These specific parameter combinations hold a distinct advantage – they offer a direct, analytically derived solution to the intricate hypergeometric functions pivotal to our methodology. This direct solution cascades to the computation of Bayes factors, an integral step that ultimately determines the threshold governing the coefficients of correlation.

As previously established and emphasized, the efficiency and effectiveness of the algorithm display a marked trend as these parameters scale. Remarkably, higher values of sample size ('n'), parameter alpha ('α'), and the number of potential features ('p') correspond to a significantly improved adaptation of the algorithm. This adaptation is precisely observed in the swift and accurate estimation of the Bayes factor threshold, a keystone element that dictates the selection of relevant predictors.

With this foundational understanding now firmly established, the next logical progression involves a robust simulation that encapsulates a myriad of artificially generated datasets. Within this extensive framework, a multitude of datasets, enriched with numerous features, shall undergo analysis. Additionally, the simulation shall span a multitude of iterations, ensuring a sufficiently large sample size for attaining highly accurate results. This strategic approach holds the promise of consistently converging towards the truth and, notably, closely approximating the characteristics of the true model in the majority of instances.

The simulation aims to explore the performance of the Fast Bayesian Variable Selection (FBVS) algorithm in scenarios where the number of predictors exceeds the number of observations. To simulate the scenario of interest, we generated synthetic 100 different datasets to ensure a comprehensive coverage of various situations. The sample size for each dataset was set at 400, and the number of potential features was fixed at 1000.

For the simulation, we assumed a true model with a known set of parameters. Specifically, the true model consisted of six predictors.

It is assumed that the explanatory variables are linear, and for these variables, we simulated values from the standardized normal distribution Normal(0,1). Additionally, we assumed that the response variable is linearly related to the above explanatory variables, resulting in a multiple linear model of the following form:

$$Y = \beta_0 + 0.70X_1 + 0.55X_2 + 1.25X_3 + X_4 + 1.8X_5 + 1.3X_6 + \epsilon \tag{3.5}$$

In this model, the first six features are considered as true predictors. Their significance is evident from the coefficients assigned to them and as it had been shown before $\epsilon \sim Normal(0, \sigma^2)$.

To create correlated predictors and response for each simulation run, we employed the Cholesky decomposition method. This method ensured that the generated predictors exhibited the specified correlation structure. By utilizing the Cholesky decomposition, we transformed a covariance matrix into a lower triangular matrix, allowing us to generate correlated predictors while preserving the desired correlation relationships. This technique is crucial in maintaining the integrity of the correlation structure across the dataset. Additionally, random noise was introduced to the response variable to create a realistic dataset for analysis. This noise factor contributes to the inherent variability seen in real-world data.

The first pivotal step of the Fast Bayesian Variable Selection (FBVS) algorithm entails the calculation of Pearson correlation coefficients between each predictor and the response variable. These correlation coefficients serve as an initial measure of potential predictor relevance. Subsequently, threshold values are established using a hypergeometric function. This threshold calculation process is designed to discern noteworthy correlations by evaluating them against the computed threshold.

The process of threshold calculation involves:

- Specifying the necessary parameters for the hypergeometric function. - Defining a custom function that encapsulates the hypergeometric equation needed for threshold computation, as outlined previously. - Employing the `uniroot` function to identify the root of the hypergeometric equation within specified tolerance levels. This resulting root value signifies the threshold for squared correlations.

Based on the threshold derived from the hypergeometric calculation, the FBVS algorithm proceeds to identify and select relevant predictors. By comparing the squared correlation coefficients of each predictor with the threshold, the algorithm distinguishes predictors with substantial correlation to the response variable. Predictors exceeding the threshold are deemed significant and are retained for further analysis, forming a subset of potentially influential variables.

This systematic approach to predictor selection is at the heart of the FBVS algorithm, enabling it to identify promising variables for model inclusion. The incorporation of both correlation analysis and thresholding ensures a rigorous methodology for predictor assessment and selection.

A prominent trend emerges, showcasing that the occurrence of precisely six features is the most frequent outcome. Notably, this outcome transpired many times during the simulations, solidifying the expected behavior. This occurrence, however, manifested only once across all simulations, indicating its rarity.

The influence of initial beta values becomes apparent when examining the distribution of feature retention. The initial betas played a pivotal role in shaping the simulation results. This variability indicates the sensitivity of this feature to different parameter settings.

Moreover, upon closer examination of the first six features, intriguing patterns emerge. The simulation data reveals the following counts for these features: 100, 93, 100, 100, and 100. Evidently, some features r were consistently retained across all 100 simulations, each boasting a perfect retention rate.

3.2.4 Adjustment in Multiple Regression with mixtures of g priors

The advantage of using the g prior over mixture methods lies in the closed-form expression for the Bayes Factor. This closed-form expression allows for efficient computation and analysis, making it useful for performing feature selection with the algorithm. By substituting specific values for a and θ, the equation provides a numerical value for the Bayes Factor, which can then be used to make informed decisions about the relevance of features in the model.

Additionally, employing the aforementioned methodology yields the value for the coefficient of determination (R^2). This enables the selection of specific values for a, allowing for further extension and refinement of the analysis.

The focus is on the comparison of two nested models which will differ in only one covariate j, the Bayes factor will be like the equation (2.7) and M_γ will be the model, from null-based strategy. The scenery will be the same with global empirical Bayes with common g to all models. So, the resulting Bayes factor is the following:

Let us consider a model M_γ with p_γ covariates.

Then, we define model $M_{\gamma_{\setminus j}}$ as the model with all covariates of M_γ except X_j for $j = 1, \ldots, p_\gamma$.

Hence, model $M_{\gamma-j}$ will contain $p_\gamma - 1$ covariates and the variable inclusion indicator of X_j will be set equal to zero, that is $\gamma_j = 0$ under $M_{\gamma-j}$. Let R^2_γ be the coefficient of determination (R-squared) for the full model (with all predictor variables) and $R^2_{\gamma-j}$ be the coefficient of determination for the reduced model (with all predictor variables except for a specific predictor of interest).

We focus our attention on the comparison of the two nested models that differ only by one covariate X_j: M_γ and $M_{\gamma-j}$. It can be noted that here M_γ is considered fixed and it can be also the full model M_F.

We can express the relative Bayesian evidence between M_γ and $M_{\gamma-j}$ as the ratio of the two null-based Bayes Factors of each competing model, that is

$$B_{M_\gamma, M_{\gamma-j}} = \frac{B_{M_\gamma,0}}{B_{M_{\gamma-j},0}}.$$

Using equation (17) of Liang et al. we can write

$$B_{M_\gamma, M_{\gamma-j}} = \frac{\frac{a-2}{p_\gamma+a-2} F_2\left(\frac{n-1}{2}, 1; \frac{p\gamma+a}{2}; R^2_\gamma\right)}{\frac{a-2}{p_\gamma-1+a-2} F_2\left(\frac{n-1}{2}, 1; \frac{p\gamma-1+a}{2}; R^2_{\gamma-j}\right)} \tag{3.6}$$

where R^2_γ and $R^2_{\gamma-j}$ are the usual coefficients of determination of the two models, respectively. This equation represents the Bayes Factor $BF_{M_\gamma, M_{\gamma_{-j}}}$, which quantifies the strength of evidence for model M_γ over the alternative model $M_{\gamma_{-j}}$, given a parameter a. This equation is used in the context of feature selection using an algorithm where a and θ are specified with certain values.

The key insight here is that the sample squared partial correlation coefficient $\rho^2_{\gamma,j}$ represents the proportion of unexplained response variability between the two models, and this relationship allows us to express the Bayes Factor formula in terms of $\rho^2_{\gamma,j}$ and R^2_γ. The partial correlation coefficient is defined as the proportion of the unexplained response variability between the two models M_γ and $M_{\gamma-j}$ with respect to the smaller one, that is:

$$\rho^2_{\gamma-j} = \frac{R^2_\gamma - R^2_{\gamma-j}}{1 - R^2_{\gamma-j}}, \tag{3.7}$$

from which

$$R^2_{\gamma-j} = 1 - \frac{1 - R^2_\gamma}{\rho^2_{\gamma-j}}. \tag{3.8}$$

This formula allows you to relate the ${R_\gamma}^2$ value for a specific predictor in the context of multiple regression to the corresponding squared partial correlation coefficient. The partial correlation coefficient provides a measure of the strength of the relationship between the predictor and the response variable after accounting for the influence of other predictors in the model. So we can insert (3.8) in (3.6), at the denominator, and solve the equation

$$B_{M_\gamma, M_{\gamma-j}} = \theta \tag{3.9}$$

with respect to $\rho^2_{\gamma-j}$ and we will have the partial correlation threshold.

In the initial step, corresponding to simple regression within the algorithm, we set a correlation threshold at 0.010. Under this criterion, the first feature, a true and vital component of our model, was consistently retained across 84 instances, while all other features were maintained 100 times across 100 different datasets. This stringent selection process ensured the retention of the essential feature, affirming its significance.

In the subsequent step, wherein two models were considered, differing solely by the inclusion or exclusion of a single feature X_j, the algorithm demonstrated remarkable precision. Specifically, the true features were accurately identified: Feature 1 was detected 100 times, Feature 2 was identified 93 times, and Features 3,4 and 5 were consistently detected 100 times each.

Conversely, features deemed non-significant were appropriately identified as such. One of these non-significant features was found 2 times, while another set of 3 features was identified 1 times, further affirming the algorithm's ability to discern and prioritize relevant components within the model.

Additionally, in the first step of the algorithm with the g prior, during the initial step, several variables were identified as true features. However, in the second and pivotal step, the results mirrored those obtained from the mixture of g priors. This consistency underscores the robustness of the algorithm's performance. In this step, a slightly higher correlation threshold of 0.020 was employed, leading to similar outcomes, thus reinforcing the reliability and accuracy of the model selection process.

Table 3.2: Bayes Factor with Feature Selection, P=1000

	Features					
n	1st	2nd	3rd	4th	5th	6th
200	16	13	20	20	20	20
300	20	19	20	20	20	20
400	19	19	20	20	20	20
500	20	20	20	20	20	20

Table 3.3: Bayes Factor with correlation thresholds, P=1000

n	Pearson Correlation	Partial Correlation
200	0.021	0.041
300	0.010	0.025
400	0.015	0.027
500	0.008	0.020

In the tables, we present simulation scenarios for different sample sizes and numbers of features, with specific parameter values set at $\alpha = 2.3$ and $\theta = 1$. These simulations are designed to explore the behavior of the model under varying conditions, shedding light on the impact of sample size and feature count on the outcomes.

Table 3.4: Bayes Factor with Feature Selection, n=400

	Features					
P	1st	2nd	3rd	4th	5th	6th
800	20	8	20	20	20	20
900	20	16	20	20	20	20
1000	20	19	20	20	20	20
1100	20	20	20	20	20	20

Table 3.5: Bayes Factor with correlation thresholds, n=400

P	Pearson Correlation	Partial Correlation
800	0.010	0.020
900	0.010	0.022
1000	0.010	0.025
1100	0.010	0.020

3.2.5 Simulation with Ozone data

In our study, we specifically utilized the Ozone dataset for our simulations. This choice was made due to our prior knowledge about the performance of certain priors and criteria, such as BIC and AIC, when applied to this specific dataset. Despite the small number of features, the correlations present in the data are grounded in factual information and are not subject to simulation errors. Moreover, this dataset has been extensively analyzed in various studies in bibliography, providing a reliable basis for our experiments.

Our simulation outcomes closely aligned with the results obtained using the BAS package, enhancing the credibility of our findings. Notably, when employing the Fast Bayesian Variable Selection (FBVS) algorithm under the g prior, the selected variables mirrored those identified using the BAS package. This consistency strengthens the reliability of our results, particularly in scenarios where these variables were consistent between the two methods.

In contrast, with the extension to hyper-g prior, there was a single variable discrepancy. However, upon closer examination, the probability of inclusion for this variable stood at 0.52, introducing an element of uncertainty. It is noteworthy that under both priors, the most probable variables were either included or excluded based on their low inclusion probabilities. This observation underscores the careful consideration and evaluation required when interpreting the results obtained through the hyper-g extension.

3.3 Algorithms for FBVS

The Fast Bayesian Variable Selection (FBVS) algorithms employed in this study, attributed to Ntzoufras and Paroli, demonstrate versatility in accommodating various scenarios of sample sizes and numbers of features. These algorithms are adept at handling a wide array of data complexities, making them invaluable tools in statistical analysis. The fundamental concept underlying these algorithms remains consistent with the original approach using the g prior. The only modification lies in the final equation, where the algorithm operates under the framework of the mixture of g priors. Despite this change, the core idea and methodology driving these algorithms remain true to the principles of the g prior, ensuring their robustness and reliability in diverse analytical contexts.

Algorithm 1 Single Covariate Screening Algorithm

```
Require: Bayes Factor threshold $\theta$; Algorithm Approach: value of $g$; sample size $n$; number of covariates under consideration $p$; response $Y$ and covariates $X_j \in V = \{X_1, \ldots, X_p\}$.
Step 1: Calculate the sample correlation coefficients for all covariates $X_j \in V$
Step 2: Identify which covariates $X_j^*$ are below the correlation thresholds given by
    A: Equation (3.4)
All non-important covariates $X_j^*$ will be denoted by the set $V^*$.
Step 3: Eliminate covariates $X_j^*$ from $V$.
Output: The final set of variables will be $X(S)_j \in V(S) = V \setminus V^*$.
We denote by $p(S) < p$ the number of covariates $X(S)_j$ remained for evaluation in $V(S)$.
```

Algorithm 2 Multiple Covariate Screening Algorithm

```
Require: Bayes Factor threshold $\theta$; Algorithm Approach: value of $g$; sample size $n$; number of covariates under consideration $p$; response $Y$ and covariates $X_j \in V = \{X_1, \ldots, X_p\}$.
Step 1: Calculate the $R^2$ of the multiple regression with all covariates $X_j \in V$ and all the sample partial correlation coefficients.
Step 2: Identify covariates $X_j^*$ with sample partial correlations lower than the corresponding thresholds given by
    A: Equation (3.9)
All non-important covariates $X_j^*$ will be denoted by the set $V^*$.
Step 3: Eliminate covariates $X_j^*$ from $V$ and denote the set as $V^*$.
Output: The final set of variables will be $X(M)_j \in V(M) = V \setminus V^*$.
We denote by $p(M) < p$ the number of covariates $X(M)_j$ remained for evaluation in $V(M)$.
```

3.4 Conclusion

In this chapter, the method of Fast Bayesian Variable Screening (FBVS) was discussed. The approach involves utilizing mixtures of g-prior specifications in both simple uni-covariate and multiple regression setups. It presents a rapid Bayesian variable selection strategy applicable to Normal regression models. The technique incorporates partial and Pearson correlation coefficient thresholds and is rooted in Bayesian principles derived from posterior model odds and Bayes factor thresholds. The key objective is to reduce the model space size and identify inconsequential covariates, ultimately leading to a condensed model space that can be subjected to more conventional and computationally demanding Bayesian variable selection techniques. Importantly, this method is adaptable and can be extended to accommodate various prior settings.

Paroli and Ntzoufras initially investigated this method, focusing on the g prior's selection due to its closed-form Bayes factor computations and precise correlation thresholds. This highly efficient screening algorithm relies on just two equations. Notably, this approach has been developed and refined in the present year (2023).

The method's foundation lies in the concept of Posterior Inclusion Probability (PIP), a valuable metric for gauging the likelihood of a covariate's inclusion in the model given the observed data. Additionally, the method employs the uni-covariate Bayes factor formula for simple regression and correlation thresholds. Correlation thresholds are set by finding specific combinations of g and Pearson correlation coefficients that yield a uni-covariate Bayes factor less than or equal to a specified threshold value.

A significant expansion to this method involves the incorporation of mixtures of g priors. This expansion builds upon the work of Liang et al. (2008)(19) and aims to enhance the algorithm's efficiency in variable selection within a Bayesian context. However, the complexity of equations and the presence of hypergeometric functions present challenges in deriving closed-form solutions for certain parameters. The `hypergeo` R package plays a vital role in handling these challenges by providing solutions based on the Hypergeometric and generalized hypergeometric functions.

The simulations revealed a consistent trend of variable retention, with around six features being the most frequently retained across the simulations. Additionally, the importance of initial beta values in shaping the simulation outcomes was evident, as well as the influence of parameter settings.

The Multiple Covariate Screening Algorithm was introduced for scenarios involving multiple predictors. This algorithm identifies relevant predictors based on their sample squared partial correlation coefficients and their relationships with the coefficient of determination. The Bayes Factor formula was derived to quantify the strength of evidence for a model over its alternative, given a parameter g for g priors and parameter a for mixture of g priors with hyper-g prior.

In the simulation experiments conducted, it was observed that the findings were largely consistent across various approaches that validate the algorithms. Specifically, when applying different methods to the same dataset, the results showed remarkable similarities. This consistency underscored the robustness of the algorithms, reaffirming their reliability and effectiveness in analyzing data.

Furthermore, in the simulation with real data, particularly focusing on the ozone dataset comprising solely numerical variables, two distinct algorithms were employed: Fast Bayesian Variable Selection under the g prior and under the mixture of g prior. Additionally, the same dataset was subjected to analysis using the toy example presented in the second chapter, employing the `BAS` package in R. In both cases, employing the g prior and the hyper g prior, the results demonstrated convergence. This agreement in outcomes further strengthens the credibility of the methodologies employed, affirming their applicability in real-world scenarios.

In conclusion, the Fast Bayesian Variable Screening method presents a valuable approach to variable selection in the context of Bayesian regression models. Its ability to efficiently screen and identify relevant predictors based on correlation thresholds and Bayes factors makes it a promising tool for tackling complex datasets.

Chapter 4

Conclusion

In this master thesis, a comprehensive exploration of Bayesian model selection and variable inclusion techniques was undertaken, with a particular focus on the context of linear models. The journey began by establishing the foundations of the normal linear model, setting the stage for the subsequent introduction to the Bayesian paradigm. A pivotal concept, the Bayes Factor, emerged as a critical tool that guided the course of this research.

The initial steps into the Bayesian realm led to a profound understanding of the model selection landscape. The investigation commenced with the examination of the g-prior framework, encapsulating the general Bayesian model selection process. Within this framework, the Bayesian Model Average (BMA) approach was introduced, a technique that amalgamates predictions from multiple models, leveraging their posterior probabilities. This innovative approach provided a more accurate estimation of the relationship between predictor variables and the response variable.

Further progression unveiled the Bayesian Adaptive Sampling (BAS) methodology. A versatile technique, BAS harnessed Bayesian model weighting through stochastic or deterministic sampling without the need for resampling from posterior distributions. BAS, by adaptively estimating marginal inclusion probabilities during sampling, addressed pivotal concerns in variable selection.

The exploration then transitioned to the Stochastic Search Variable Selection (SSVS) method, an ingenious approach introduced by George and McCulloch(14). This method treated variable inclusion as a Bernoulli trial, utilizing a mixed prior to represent the potential presence or absence of each variable. Initially conceived in linear models, SSVS found its applications expanding to diverse models, exemplified by log-linear models for multi-way contingency tables.

The research also delved into the Kuo & Mallick (KM) Sampler(18), a model that exhibited distinctive structural characteristics when compared to SSVS. The selection process of prior specifications set KM apart. Similarly, the Gibbs Variable Selection (GVS) method, pioneered by Dellaportas et al.(8), harnessed indicator variables as components of model equations, presenting an alternative avenue in model selection.

The study further drew from Liang et al.'s foundational work (2008)(19) to explore the paradoxes and modifications of the g-prior under various contexts. In particular, the incorporation of mixture models unveiled a realm of possibilities, culminating in the development of the FBVS algorithm. This algorithm's efficacy and benefits were expounded upon, showcasing its precision and robustness.

A meticulous analysis of the FBVS algorithm revealed its adaptability in various scenarios. The derivation of essential equations facilitated precise results. Additionally, correlation thresholds were elucidated, their mathematical representation proving invaluable in practice.

The algorithm's reach extended to mixtures of the g-prior. As explored through the lens of Liang's equation, challenges emerged that required innovative problem-solving. The methodology proposed here overcame these challenges, reflecting its resilience and applicability.

Empirical validation through simulations reaffirmed the algorithm's effectiveness. The impact of sample size on accuracy was noted, demonstrating an upward trend as the sample size increased. A simulation-based approach showcased the algorithm's prowess in scenarios with a high number of potential features.

In summation, this master thesis navigated the intricate landscape of Bayesian model selection techniques. By probing various methodologies, algorithms, and modifications, a comprehensive understanding emerged, unveiling their underlying mechanisms, strengths, and limitations. The thesis contributes to the broader discourse on Bayesian methods and paves the way for continued advancements in model selection and variable inclusion.

4.1 Future Work

Despite the promising results showcased by the Fast Bayesian Variable Selection (FBVS) algorithm, a crucial limitation arises in its current formulation. Specifically, in the context of multiple regression, the algorithm introduces a hypergeometric function to establish threshold values for feature inclusion. However, a gap emerges as the coefficient of determination remains unincorporated within the same framework as observed in the multiple regression under g-prior. This discrepancy underscores the need for further exploration and refinement.

In essence, future pursuits can be directed towards refining the algorithmic architecture to encompass the coefficient of determination and its integration within the hypergeometric function. By closing the existing gap and establishing a unified formula, this approach holds the potential to bolster the algorithm's performance, furnishing a more resilient solution to the challenges posed by high-dimensional data scenarios.

Moreover, within the realm of Bayesian regression analysis, a spectrum of other priors beckons exploration. The Generalized g-Prior, introduced by Maruyama and George (2011)(23), extends the conventional g-prior to enable varying prior variances for individual regression coefficients. This adaptability accommodates prior information about the variability of distinct predictors, contributing to a more nuanced modeling process. Similarly, the Robust Prior, proposed by Bayarri et al. (2012)(3), addresses the impact of outliers and influential observations in regression modeling. By assigning diminished weight to extreme data points, this prior fosters robust parameter estimation, particularly when data anomalies challenge distributional assumptions.

These diverse priors provide researchers with a spectrum of options to incorporate prior information and assumptions into Bayesian regression analyses. The choice of prior hinges on data characteristics, research objectives, and underlying assumptions. Each prior offers distinct advantages and considerations, underscoring the importance of tailored selection aligned with the research context.

Appendix A

Inverse Gamma

The main function of the inverse gamma distribution is in Bayesian probability, where it is used as a marginal posterior (a way to summarize uncertain quantities) or as a conjugate prior (a prior is a probability distribution that represents your beliefs about a quantity, without taking any evidence into account). In other words, it's used to model uncertain quantities.

In Bayesian statistics, the inverse gamma distribution often serves as a prior distribution for the precision (inverse of the variance) of a normal distribution. If X follows a normal distribution with mean μ and precision φ, then the prior distribution for φ is commonly chosen as:

$$\varphi \sim \text{Inverse Gamma}(\alpha, \beta)$$

This choice is particularly common in Gaussian models, where it acts as a conjugate prior, simplifying Bayesian computations.

Probability Density Function (PDF):

$$f(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot x^{-\alpha-1} \cdot e^{-\frac{\beta}{x}}$$

A.1 Spike-and-Slab Prior

In Bayesian statistics, a *spike-and-slab prior* is a type of hierarchical prior that combines a point mass (spike) with a continuous distribution (slab). This prior is often used in variable selection and model averaging.

The spike-and-slab prior is typically expressed as a mixture distribution:

$$\pi(\theta) = \pi_{\text{spike}} \cdot \delta_0(\theta) + \pi_{\text{slab}} \cdot f(\theta)$$

where:

- π_{spike} is the weight assigned to the spike component,
- $\delta_0(\theta)$ is a Dirac delta function at zero (point mass),
- π_{slab} is the weight assigned to the slab component,
- $f(\theta)$ is a continuous density function representing the slab.

The spike-and-slab prior is useful for modeling scenarios where some model parameters are exactly zero with high probability (spike), while others are allowed to take non-zero values (slab).

Bibliography

[1] Abramowitz, Milton, and Irene A. Stegun (1970). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, **55**, US Government printing office.

[2] Barbieri, Maria Maddalena, and James O. Berger (2004). Optimal Predictive Model Selection. *Annals of Statistics* **32**(3), 870-897

[3] MJ Bayarri, JO Berger, A Forte, and G Garcıa-Donato (2012). Criteria Bayesian model choice with application to variable selection. *The Annals of Statistics*, **40**(3), 1550–1577.

[4] Maurice S Bartlett (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, **44**(3-4), 533–534.

[5] Philip J Brown, Marina Vannucci, and Tom Fearn (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**(3), 627–641.

[6] Merlise A Clyde, Joyee Ghosh, and Michael L Littman (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, **20**(1), 80–101.

[7] Wen Cui and Edward I George (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, **138**(4), 888–900.

[8] Petros Dellaportas, Jonathan J Forster, and Ioannis Ntzoufras (2000). Bayesian variable selection using the Gibbs sampler. *Biostatistics-Basel-*, **5**, 273–286.

[9] Petros Dellaportas, Jonathan J Forster, and Ioannis Ntzoufras (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**(1), 27–36.

[10] Carmen Fernandez, Eduardo Ley, and Mark FJ Steel (2001). Benchmark Priors for Bayesian Model Averaging. *Journal of Econometrics*, **100**(2), 381–427.

[11] Dean P Foster and Edward I George (1994). The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics*, **22**(4), 1947–1975.

[12] Edward I George and Dean P Foster (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika*, **87**(4), 731–747.

[13] Edward I George and Robert E McCulloch (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**(423), 881–889.

[14] Edward I George and Robert E McCulloch (1995). Stochastic Search Variable Selection. *Markov Chain Monte Carlo in Practice*, **68**(1), 203–214.

[15] Edward I George and Robert E McCulloch (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, 339–373.

[16] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky (1999). Bayesian Model Averaging: A Tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors. *Statistical Science*, **14**(4), 382–417.

[17] Robert E Kass and Adrian E Raftery (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

[18] Lynn Kuo and Bani Mallick (1998). Variable Selection for Regression Models. *Sankhya: The Indian Journal of Statistics, Series B*, 65–81.

[19] Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, **103**(481), 410–423.

[20] Dennis V Lindley (1957). A Statistical Paradox. *Biometrika*, **44**(1/2), 187–192.

[21] Anastasia Lykou and Ioannis Ntzoufras (2013). On Bayesian Lasso Variable Selection and the Specification of the Shrinkage Parameter. *Statistics and Computing*, **23**, 361–390.

[22] David Madigan and Adrian E Raftery (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, **89**(428), 1535–1546.

[23] Yuzo Maruyama and Edward I George (2011). Fully Bayes Factors with a Generalized G-Prior. *The Annals of Statistics*, **39**(5), 2740–2765.

[24] Ioannis Ntzoufras (2011). *Bayesian Modeling using WinBUGS*, **698**. John Wiley & Sons.

[25] Ioannis Ntzoufras, Jonathan J Forster, and Petros Dellaportas (2000). Stochastic Search Variable Selection for Log-Linear Models. *Journal of Statistical Computation and Simulation*, **68**(1), 23–37.

[26] Ioannis Ntzoufras, Roberta Paroli, et al. (2021). Bayesian Screening of Covariates in Linear Regression Models Using Correlation Thresholds. In *BOOK OF SHORT PAPERS–SIS2021*, Pearson, 1232–1237.

[27] RB O'Hara and MJ Sillanpää (2009). A Review of Bayesian Variable Selection Methods: What, How, and Which. *Bayesian Analysis*, **4**(1), 85–118.

[28] William E Strawderman (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *The Annals of Mathematical Statistics*, **42**(1), 385–388.

[29] Luke Tierney and Joseph B Kadane (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, **81**(393), 82–86.

[30] Arnold Zellner (1986). Bayesian Estimation and Prediction Using Asymmetric Loss Functions. *Journal of the American Statistical Association*, **81**(394), 446–451.

[31] Arnold Zellner and Aloysius Siow (1980). Posterior Odds Ratios for Selected Regression Hypotheses. *Trabajos de estadística y de investigación operativa*, **31**, 585–603.