Model Uncertainty in Modelling the Growth of Countries

by

Moez Abeidi
(ID: 1814920)

Supervised by: Professor Mark F.J. Steel

WARWICK
THE UNIVERSITY OF WARWICK

ST415: Statistics Masters Dissertation

Department of Statistics
University of Warwick
United Kingdom

May 2022

Abstract

This project sets out to tackle the question: “What is the true uncertainty behind picking this model?”. We focus on the context of economic growth, however model uncertainty is an important phenomenon to be aware of in many different contexts, as it appears in almost all situations where there is uncertainty in variable selection. We take particular focus on Bayesian Model Averaging, and how this method compares to model selection and other model averaging methods. We use and compare these approaches to real data used in the related literature, and formulate our conclusions based on the results. In addition, we explore situations where the set of all possible models is too large to be handled by our modern CPUs and how BMA deals with this.

Acknowledgements

I would like to thank my supervisor Professor Mark F.J. Steel for his guidance and support throughout the project. In addition, I would like to thank the Department of Statistics at the University of Warwick for their support and organisation of the dissertation module.

Acronyms

In this project, several acronyms are used. To facilitate understanding, below is a table of the common acronyms used throughout this document and in the related literature.

BMA	Bayesian Model Averaging
FMA	Frequentist Model Averaging
BMS	Bayesian Model Selection
PMP	Posterior Model Probabilities
PIP	Posterior Inclusion Probabilties
MCMC	Markov Chain Monte Carlo
LPS	Log Predictive Score
iid	Independently and Identically Distributed
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
MMA	Mallow's Model Averaging
JMA	Jackknife Model Averaging
WALS	Weighted Average Least-Squares
BACE	Bayesian Averaging of Classical Estimates
OLS	Ordinary Least Squares
LASSO	Least Absolute Shrinkage and Selection Operator
GLM	Generalised Linear Model
PDF	Probability Density Function
PMF	Probability Mass Function
MLE	Maximum Likelihood Estimator
GDP	Gross Domestic Product

Notation

To facilitate understanding, below is a table of the most common notation used throughout this document and their corresponding interpretations - unless stated otherwise. Note the notation is generally similar to what is used in the literature, with a few differences for consistency.

y	The observations from the data
n	Number of data observations
β	The vector of model parameters
α	The common intercept
σ^2	The model variance
$\mathcal{M}$	Model space (set of all possible covariate combinations)
M_j	Model j from $\mathcal{M}$
θ	General symbol for model parameters
γ	Probabilitiy of regressor inclusion
m	Prior mean model size
$l_y(M_j)$	Marginal Likelihood for model M_j
k	Number of regressors in dataset
ω	FMA model weights
J	Number of possible models to be considered - if unrestricted $J = \|\mathcal{M}\| = 2^k$
Z_j	Design matrix for model M_j
k_j	Number of regressors for model M_j
β_j	The vector of parameters for model M_j
q	Number of new data points
δ	The shrinkage factor

This is not a comprehensive list, and many are to be defined throughout the document.

Contents

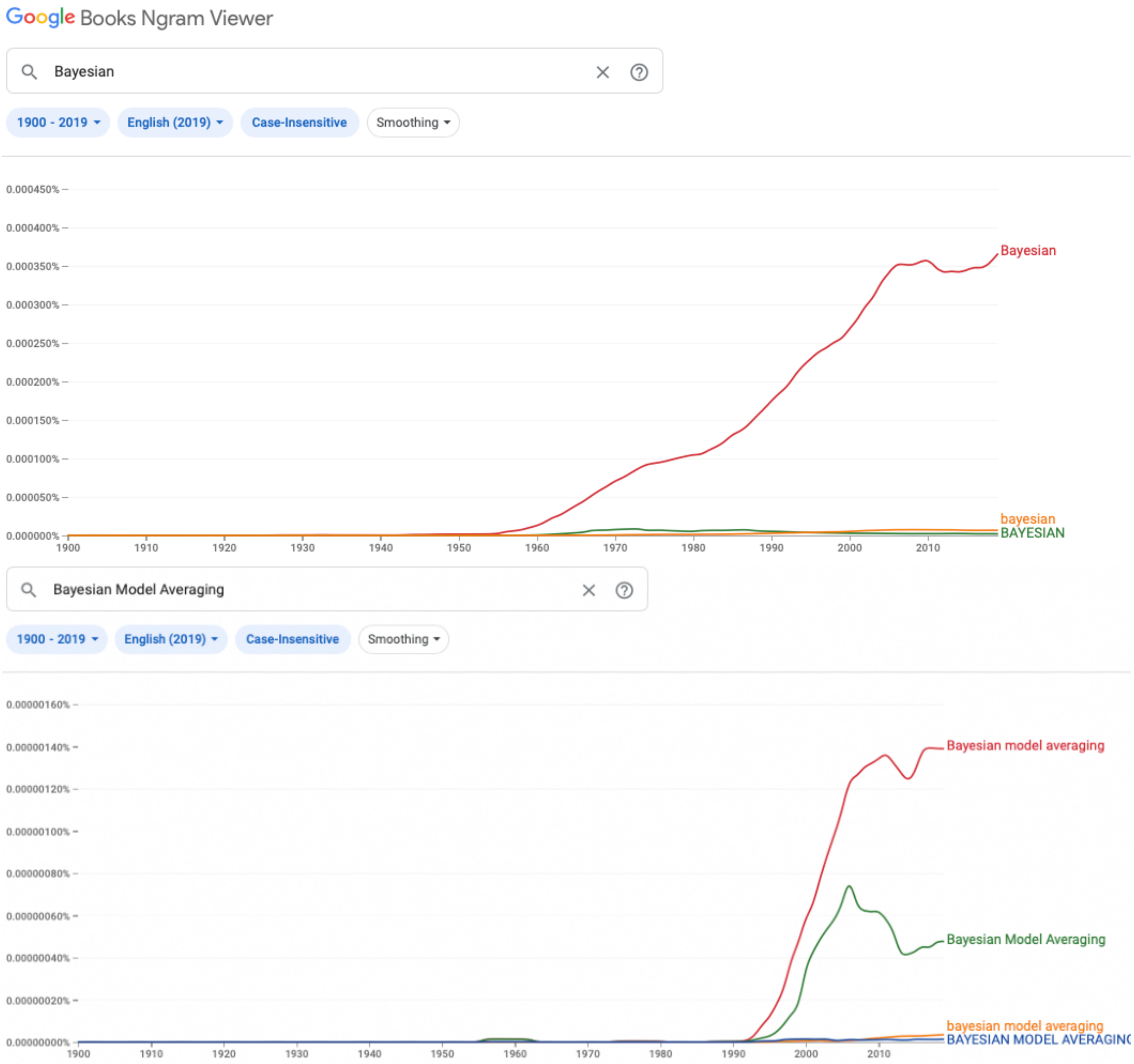
1 Introduction

The importance of model uncertainty in economic modelling should not be underestimated. Empirical work in economics will almost always be subject to large amounts of uncertainty on model specifications. A lot of this uncertainty can be attributed to the open-endness of empirical growth theory, which entails that alternative theories may suggest additional determinants of growth without necessarily excluding determinants proposed by other theories [Ley and Steel, 2007]. At the theoretical level, there is an absence of such trade-off, which ultimately leads to substantial model uncertainty over which variables should be included in a growth regression. According to [Steel, 2020], this model uncertainty may be the consequence of the existence of many different economic theories and the manners in which they can be implemented in empirical models, or perhaps due to other aspects such as assumptions about the independence or heterogeneity of the variables. It is important to be aware of such uncertainties.

Model averaging is a method that has been employed in the literature to attempt to tackle such uncertainty. Using model averaging, we take into consideration information from all models in the model space to come up with a final model. There are various different model averaging methods, however they can roughly be split into Bayesian Model Averaging (BMA) and Frequentist Model Averaging (FMA), with a few methods straddling the line between the two.

In contrast to model averaging is model selection, in which the best performing model under a certain criterion (AIC or BIC for example) is chosen, or is chosen subjectively. Information from all other models is ignored no matter how well they perform, and all inference is made from the one model. Model selection is thus conditioning on the chosen model, without regard to the aspects of reality captured in the other models, which leads to an underestimation of the true uncertainty. Selecting the one best model can often ignore important information in other models, especially when there is no "true" model - as is often the case in economics [Steel, 2020].

Bayesian Model Averaging, and Bayesian statistics in general, are seeing a large growth, with more and more scientific papers using Bayesian methods as opposed to frequentist methods [Hackenburger 2019]. This is in part due to the development of modern Markov Chain Monte Carlo (MCMC) methods that aid in Bayesian inference, but also due to the flexibility Bayesian methods offer - such as providing a natural framework for model averaging. Using BMA, we can also obtain all the benefits of Bayesian methods, such as obtaining marginal densities for the variables, and predictive densities for future observations. In addition, BMA has been shown to predict better than model selection methods whilst also addressing the issue of covariate collinearity [Fernandez et al., 2001b]. Figure 1 highlights the growth in BMA and Bayesian statistics in the last few decades, clearly indicating an increase in frequency in the use of these terms in the last 120 years.



Bayesian

Bayesian Model Averaging

Figure 1: Google Ngram results for search terms ”Bayesian” and ”Bayesian Model Averaging”.

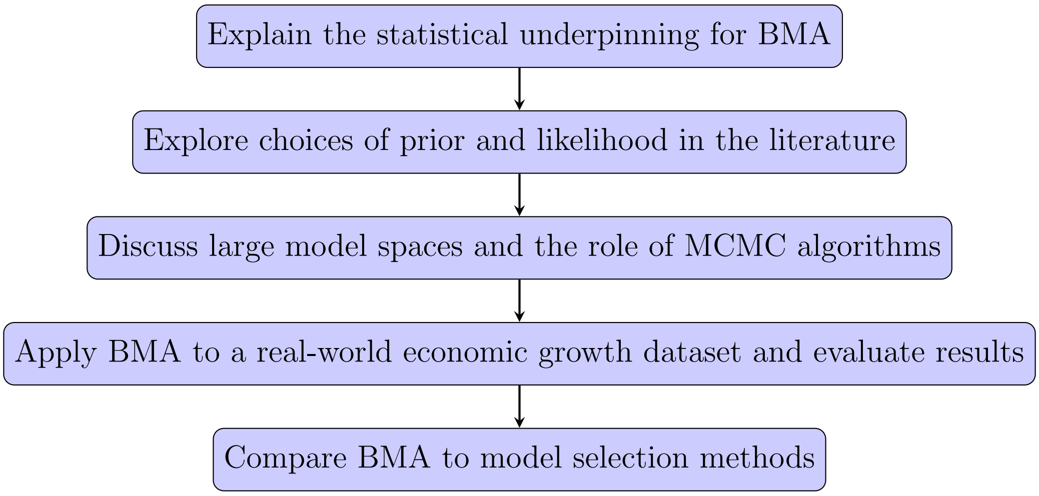
As is required with Bayesian statistics, a prior and likelihood must be specified. In this project, we will explore the various choices of priors in the literature, both on the model space (the set of all possible covariate combinations), and on the parameters.

For BMA, we will perform Bayes' rule over both the parameter space and the model space, ensuring we have the building blocks to perform model averaging. As shown in [Ley and Steel, 2009], certain prior assumptions can have a substantial effect on posterior results. Various prior settings suggested in the literature can give considerable differences in the posterior, thus it is important to be aware of this prior sensitivity. In this project, the theoretical reasoning behind the variation of the posterior given different priors is explained.

We then apply BMA to a real-world economic growth dataset and discuss the results.

In the growth regression literature, the normal linear regression model is widely used, however there have been important developments in other model settings, in particular Generalised Linear Models (GLMs). In this project, we take a focus on the normal linear model as this is the most widely used likelihood function and potentially the most widely applicable; however, we also briefly consider the GLM case.

In brief, this project aims to:



Explain the statistical underpinning for BMA

Explore choices of prior and likelihood in the literature

Discuss large model spaces and the role of MCMC algorithms

Apply BMA to a real-world economic growth dataset and evaluate results

Compare BMA to model selection methods

Despite being primarily focused on economic growth, this project has applications in a wide domain of fields. In essence, the methods discussed can be applied to any discipline that deals with model uncertainty in variable selection, such as: the social sciences, medicine, biology, business, data science, machine learning, etc.

2 Bayesian Model Averaging

2.1 Basics of Bayesian Inference

Before defining Bayesian Model Averaging, it is important to define the basics of Bayesian statistics. Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability distribution for a hypothesis as more information becomes available.

First, we define the main components of Bayes' theorem:

The prior $p_{\boldsymbol{\theta}}(\theta)$ is the information we have on the parameter $\boldsymbol{\theta}$, in particular our assumption on its distribution, *prior* to testing this assumption against the data.

The likelihood $p_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta)$ is a function that describes the probability of the observed data as a function of the parameters of the chosen statistical model. A likelihood function takes the dataset as a given, and represents the likeliness of different parameters for your distribution. Think of it as the function that gives us an idea of "how well" the data summarises the parameters [Glen, 2021].

We are now ready to define the posterior distribution in Bayes' theorem.

Theorem 1. (Bayes' rule) [Houssineau, 2021]
Let the prior distribution of $\boldsymbol{\theta}$ on Θ be $p_{\boldsymbol{\theta}}(\cdot)$ and let the random variable $\boldsymbol{y}$ on $\mathcal{Y}$ have conditional distribution $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\cdot|\cdot)$. If y is a given realisation of $\boldsymbol{y}$, then the posterior distribution of $\boldsymbol{\theta}$ given the observation y is

$$p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta|y) = \frac{p_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta)p_{\boldsymbol{\theta}}(\theta)}{p_{\boldsymbol{y}}(y)}, \tag{1}$$

where $p_{\boldsymbol{y}}(\cdot)$ is the marginal distribution of $\boldsymbol{y}$, defined as:

$$p_{\boldsymbol{y}}(y) = \int_{\Theta} p_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta)p_{\boldsymbol{\theta}}(\theta)d\theta. \tag{2}$$

See Figure 2 for a graphical representation.

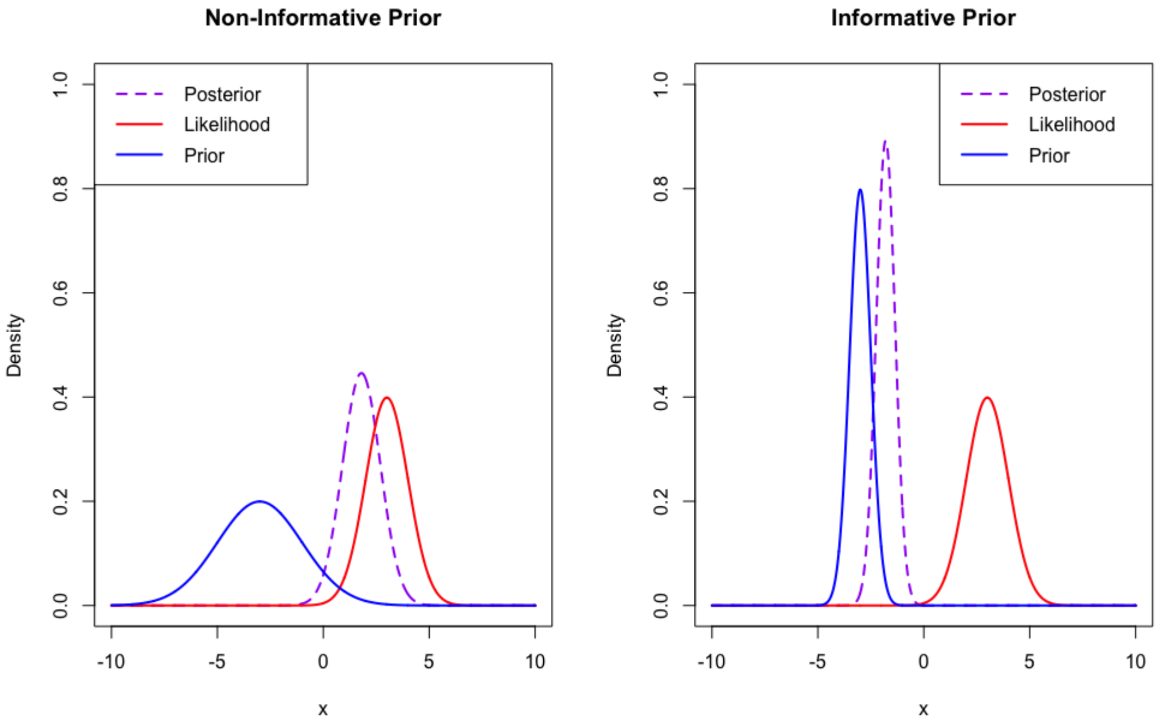


Figure 2: Bayes' Rule Illustrated

Figure 2 shows that the posterior probability $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta|y)$ is based on a combination of prior beliefs $p_{\boldsymbol{\theta}}(\theta)$ and the likelihood $p_{\boldsymbol{y}|\boldsymbol{\theta}}(y|\theta)$, normalised by the marginal distribution of $\boldsymbol{y}$, $p_{\boldsymbol{y}}(y)$.

In the left panel is a prior distribution that we are not "sure about" (high variance), whereas in the right panel is a prior distribution we are "sure about" (low variance).

This formula shall be used on the parameters θ, but also on the model space $\mathcal{M}$, where M_j, for $j = 1, ..., J$, is the model selected from the model space.

2.2 Basics of Bayesian Model Averaging

The idea behind BMA is to incorporate the information from the other models in the model space $\mathcal{M}$. This is done by allocating weights for models (based on "merit") and averaging a quantity of interest Δ, over its values in the other models.

Luckily, the choice of weights in BMA is very natural and is in fact the posterior probabilities of the model. As we will see later, other model averaging methods may not have such a natural choice of weights.

First, we need to define the model posterior probabilities. A formal Bayesian approach is to treat the model index as a random variable and use the data to conduct inference on it. We thus consider $M_j, j = 1, ..., J$ (each model is a unique combination of covariates) for the J models in the model space $\mathcal{M}$.

From Theorem 1, Bayes' rule states:

$$P(M_j|y) \propto p(y|M_j)P(M_j). \tag{3}$$

Now, we consider the prior $P(M_j)$, which is explored in section 2.3.2, on $\mathcal{M}$ and the likelihood distribution $p(y|M_j)$ - which also can be denoted as $l_y(M_j)$, the marginal likelihood of M_j. Note throughout the rest of this project, we shall use $l_y(M_j)$, as is most common in the literature.

By the Law of Total Probability, the marginal likelihood is simply the model likelihood integrated with the prior on the parameters of M_j, denoted by $p(\theta_j|M_j)$.

$$\therefore l_y(M_j) = \int p(y|\theta_j, M_j)p(\theta_j|M_j)d\theta_j. \tag{4}$$

Thus, BMA is performed as such:

$$P_{\Delta|y} = \sum_{j=1}^{J} P_{\Delta|y,M_j} P(M_j|y), \tag{5}$$

where Δ is not a model-specific quantity.

The formula tells us that the posterior of the quantity Δ given a model M_j, is averaged over all its values in the other models M_j, for $j = 1, ..., J$, using the posterior model probability (PMP) of M_j as weights. This implies a fully probabilistic treatment of the model uncertainty, just like the (well-known) parameter uncertainty.

2.3 The Bayesian Model

2.3.1 Priors Over Model Parameters

Bayesian inference is performed over the model parameters to determine the coefficients for the variables in the model. This simply corresponds to Bayesian regression conditioned on a model. Of course, for BMA, the components of Bayes' rule are theoretically defined for all models in the model space $\mathcal{M}$.

In keeping with the growth regression literature, for this project we shall consider primarily the normal linear regression model, and following the analyses of [Levine and Renelt, 1992], [Sala-i Martin, 1997], [Fernandez et al., 2001b] and [Ley and Steel, 2009], we consider GDP growth for n countries (grouped in matrix y), regressed against k explanatory variables and an intercept α.

Likelihood 1. (Normal Linear Regression) The normal linear regression model for M_j is defined as:

$$y|\alpha, \beta_j, \sigma, M_j \sim N(\alpha\iota_n + Z_j\beta_j, \sigma^2 I), \tag{6}$$

where Z_j is the matrix of the $0 \leq k_j \leq k$ variables, ι_n is a vector of n rows of ones, $\beta_j \in \mathbb{R}^{k_j}$ the matrix of coefficients for model M_j and $\sigma \in \mathbb{R}^+$ is a scale parameter.

If we allow for any subset of regressors to be included in the model, this will give a model space size of 2^k potential models. [Levine and Renelt, 1992] and [Sala-i Martin, 1997] restrict the model to always include certain key variables, with [Levine and Renelt, 1992] allowing up to four added regressors and [Sala-i Martin, 1997] allowing exactly four added regressors. Although it may seem that taking this approach is more suitable, BMA can often deal with large model spaces using efficient MCMC methods (see section 3). This suggests that, in practice, we do not need to limit our model space size as one may think.

For the priors on the parameters θ in a given model, Zellner's g-prior introduced in [Zellner, 1986] is often used on the regression coefficients β_j. In addition, Jeffreys-style "non-informative" improper priors are often used on the parameters that are common to all models - intercept α and scale σ.

Prior 1. (Parameter Prior Structure) The prior densities are defined as such

$$p(\beta_j|\alpha, \sigma, M_j) = f_N^{k_j}(\beta_j|0, \sigma^2 g(Z_j' Z_j)^{-1}),$$

$$p(\alpha, \sigma) \propto \sigma^{-1},$$

thus

$$p(\alpha, \beta_j, \sigma|M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j|0, \sigma^2 g(Z_j' Z_j)^{-1}), \tag{7}$$

where $f_N^q(x|\mu, \Sigma)$ denotes the density function of a q-dimensional normal distribution, on random variable x, with mean μ and covariance matrix Σ. This prior structure is shared by many papers in the literature on covariate selection.

In essence, the prior on the parameters is a multivariate normal distribution centered on mean 0 for each variable, with the covariance to be scaled based on value of g. The covariates not represented in model M_j have prior point mass at 0. A proper prior on β_j is needed, as an improper prior would not lead to meaningful Bayes factors [Ley and Steel, 2009]. The amount of prior information requested is therefore limited to the value $g > 0$. It thus remains to define g.

Picking arbitrary g may have unintended consequences for posterior inference, however [Fernandez et al., 2001a] conclude that $g = \max\{n, k^2\}$ leads to reasonable results. They term this the benchmark prior. The choice $g = n$ was used in [Kass and Wasserman, 1995] and [Zellner and Siow, 1980] (it can be seen as corresponding to the amount of information contained in one observation) and the choice $g = k^2$ was used in [Foster and George, 1994] for their risk inflation criterion prior. In effect, in growth regression where the number of potential covariates is large, for the benchmark prior we usually take $g = k^2$ as we typically have $k^2 \gg n$. See section 2.5 for the different effects of both prior choices on posterior probabilities.

Some consider taking g as random – i.e. assigning a hyper-prior on g. This is a natural Bayesian approach to uncertainty in the value of g, and was explored further in [Liang et al., 2008]. In particular, they introduce hyper-g priors which correspond to the following family of priors:

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}, \tag{8}$$

where $a > 2$ to have a proper distribution for $g > 0$.

[Ley and Steel, 2012] consider the shrinkage factor $\delta = g/(g+1)$, and place a Beta(b, c) on it. According to [Ley and Steel, 2012], a relatively large number of priors in the literature imply a Beta prior distribution for the shrinkage factor. From equation (6) in their paper, this induces the following prior on g:

$$p(g) = \frac{\Gamma(b+c)}{\Gamma(b)\Gamma(c)} g^{b-1}(g+1)^{-(b+c)}. \tag{9}$$

[Ley and Steel, 2012] propose a benchmark beta prior by considering the beta shrinkage prior in (9), with mean shrinkage equal to the benchmark prior defined in [Fernandez et al., 2001a], and the second parameter chosen to ensure a reasonable prior variance. They recommend using $b = c \cdot \max\{n, k^2\}$ and $c = 0.01$.

An advantage to the prior structure defined in (7), for the normal linear likelihood in (6), is that the marginal likelihood for each model and Bayes' factor between two models, can be calculated in closed form [Fernandez et al., 2001a]. In particular, the posterior results for the model parameters can be calculated analytically as such:

$$p(\beta_j|\alpha, \sigma, M_j) = f_N^{k_j}(\beta_j|\delta(Z_j'Z_j)^{-1}Z_j'y, \sigma^2\delta(Z_j'Z_j)^{-1})$$

$$p(\alpha|\sigma, M_j) = f_N^1(\alpha|\bar{y}, \tfrac{\sigma^2}{n})$$

$$p(\sigma^{-2}|M_j) = f_{Ga}(\sigma^{-2}|\tfrac{n-1}{2}, \tfrac{s_\delta}{2}),$$

where $\delta = g/(g+1)$ $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $s_\delta = [\delta y' Q_{X_j} y + (1-\delta)(y - \bar{y}\iota_n)'(y - \bar{y}\iota_n)]$, for $Q_{X_j} = I_n - X_j(X_j'X_j)^{-1}X_j'$, and where $X_j = (\iota_n : Z_j)$ is of full rank. Furthermore, $f_{Ga}(\cdot|a, b)$ represents the density function of a Gamma distribution with mean a/b. The conditional independence between β_j and α, given σ, is a consequence of subtracting the regressors from their means (demeaning).

Using this prior-likelihood setting, we can compute an analytical form of the marginal likelihood by integrating out the model parameters, as in (4), as follows:

$$l_y(M_j) = p(y|M_j) \propto (g+1)^{\frac{n-1-k_j}{2}} \left[1 + g(1 - R_j^2)\right]^{-\frac{n-1}{2}}. \tag{10}$$

In addition, for each model M_j, the marginal posterior distribution of the regression coefficients β_j is a k_j-variate Student-t distribution with $n-1$ degrees of freedom, location $\delta(Z_j'Z_j)^{-1}Z_j'y$ (which is the mean if $n > 2$), scale matrix $\delta s_\delta(Z_j'Z_j)^{-1}$, and variance $\frac{\delta s_\delta}{n-3}(Z_j'Z_j)^{-1}$ if $n > 3$ [Steel, 2020].

Generally, the hierarchical prior on g would mean the marginal likelihood of a given model is not analytically known, however it would be given by the integral of (10) integrated with respect to the prior of g. This could then be sampled using MCMC methods (see section 3).

The g-prior is widely seen in the literature, and is thus relatively well understood. It has convenient properties, such as invariance under translation and rescaling of covariates, and automatic adaptation to settings with near-collinearity between covariates [Robert et al., 2007, p. 193].

2.3.2 Priors Over Models

The prior $P(M_j)$ on the model space is usually constructed from the inclusion probability of each covariate. We assume the probability γ, of each covariate entering the model is identically and independently distributed (iid). Thus,

$$P(M_j) = \gamma^{k_j}(1-\gamma)^{k-k_j}. \tag{11}$$

Note $\gamma = 0.5$ can be seen as the benchmark choice. Which gives our prior beliefs of the model M_j to be $P(M_j) = 2^{-k}$ and expected model size of $k/2$. This fixed γ value was used in [Fernandez et al., 2001b] and [Raftery et al., 1997]. For $\gamma > 0.5$, the prior setting will favour larger models, and for $\gamma < 0.5$, the prior setting will favour smaller models. This is quantified using prior odds as to be seen in section 2.4.

[Ley and Steel, 2009] explore different choices for γ, and they conclude a hyper-prior on γ is the most suitable for growth regression contexts. They argue a fixed prior on the value γ is too strong an assumption for most datasets, and that a fixed γ should not be used as a "non-informative" prior.

Many authors in the literature such as [Brown et al., 1998], [Clyde and George, 2004] and [Ley and Steel, 2009] have suggested assigning a Beta(a, b) hyper-prior on γ. This gives a prior on M_j as defined below.

Prior 2. (Model Prior Structure) The prior on the model space is defined as such

$$P(M_j) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+k_j)\Gamma(b+k-k_j)}{\Gamma(a+b+k)}, \tag{12}$$

and a Binomial-Beta prior model size of

$$P(W=w) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+k)} \binom{k}{w} \Gamma(a+w)\Gamma(b+k-w), \tag{13}$$

where we denote W as the model size.

The expected model size and model size variance is thus:

$$\mathbb{E}[W] = \frac{a}{a+b}k,$$

$$\text{Var}[W] = \frac{ab(a+b+k)}{(a+b)^2(a+b+1)}k.$$

[Ley and Steel, 2009] recommend fixing $a = 1$, and defining $b = (k-m)/m$, where m is the prior mean model size. Thus, only a suitable m needs to be specified. Picking m would be intuitively equivalent to picking a fixed γ, with $m = \gamma k$. They argue this allows for a wide range of prior behaviour and generally leads to reasonable prior assumptions, whilst facilitating prior elicitation.

Using $a = 1$ and $b = (k-m)/m$, we obtain an expected model size and model size variance as such:

$$\mathbb{E}[W] = m, \tag{14}$$

$$\text{Var}[W] = \frac{(k-m)(1+m)}{k+m}m. \tag{15}$$

The variance of W is roughly equal to m^2, when $k \gg m$.

For any $m < k/2$ the prior mode for γ will be at 0, which suggests a conservative prior stance as most prior mass is on the null model. Intuitively, this means the "evidence" in the data is responsible for the inclusion of the regressors.

2.4 Comparing Models and Evaluating Performance

One of the most important values in comparing two competing models is the Bayes' factor. In essence, the Bayes' factor is the ratio of the marginal likelihoods of the two models, and this is used to determine which model best fits the data.

Definition 1.(Bayes' Factor) The Bayes' factor for model M_i vs M_j is defined as

$$B_{ij} = \frac{l_y(M_i)}{l_y(M_j)}, \tag{16}$$

where $l_y(M_i)$ and $l_y(M_j)$ are the marginal likelihoods of models M_i and M_j respectively.

A small Bayes' factor, typically $\log(B_{ij}) < -1$, indicates evidence against M_i when compared to M_j, and $\log(B_{ij}) < -2$ indicates strong evidence [Houssineau, 2021]. Note that values less than 1 (or 0 when logs taken) favour the model on the denominator - as it has a larger likelihood.

Also note that the value of the Bayes' factor should be primarily used for comparison purposes, as its value depends on which model is compared against which other model; for instance, the Bayes' factor for M_j vs. M_i is simply $B_{ji} = 1/B_{ij}$.

Another important value to be aware of in model comparison is the prior odds.

Definition 2.(Prior Odds) The prior odds for model M_i vs M_j is defined as:

$$\frac{P(M_i)}{P(M_j)}. \tag{17}$$

The prior odds measures the ratio of prior probability assigned to model M_i and model M_j.

Definition 3.(Posterior Odds) The posterior odds for model M_i vs M_j is defined as:

$$\frac{P(M_i|y)}{P(M_j|y)} = \frac{l_y(M_i)P(M_i)}{l_y(M_j)P(M_j)}. \tag{18}$$

Thus it is evident the posterior odds between two models is the product of their Bayes' factor and their prior odds.

Posterior odds will be equal to 1 if the Bayes' factor equals the inverse of the prior odds, thus if:

$$\frac{l_y(M_i)}{l_y(M_j)} = \frac{P(M_j)}{P(M_i)}.$$

We can think of this as the data evidence required to compensate for the difference of prior assumptions [Ley and Steel, 2009].

When evaluating model performance to the test data, a scoring criterion is needed to give us a sense of how well the model performs in prediction. We shall use the Log Predictive Score (LPS), introduced by [Good, 1952], which will help us score our model performance later in section 5. An advantage of the LPS is that it is a strictly proper scoring rule, as it induces the forecaster to be honest in divulging their predictive distribution [Fernandez et al., 2001b].

In the case where we split our observations into n observations used as training data (the observations used to build the model), and q observations used as test data (the observations used to test the model performance), we use the Log Predictive Score on the q observations. Before defining the Log Predictive Score, it is important to define the predictive distribution first. The predictive distribution is the distribution of possible unobserved values conditioned on the observed values.

Definition 4. (Predictive Distribution) The predictive distribution is defined as:

$$p(y_f|y) = \int_{\Theta_j} p(y_f|y, \theta_j, M_j)p(\theta_j|y, M_j)d\theta_j, \tag{19}$$

and defined in the BMA framework as:

$$p(y_f|y) = \sum_{j=1}^{J} \left[\int_{\Theta_j} p(y_f|y, \theta_j, M_j)p(\theta_j|y, M_j)d\theta_j \right] P(M_j|y). \tag{20}$$

According to [Fernandez et al., 2001b], for $J = 2^k$ models considered, the predictive distribution given by the prior (7), and the normal linear likelihood, is expressed as:

$$p(y_f|y) = \sum_{j=1}^{2^k} \left[f_S(y_f \mid n-1, \bar{y} + \delta z'_{f,j}\hat{\beta}_j, \frac{n-1}{d_j^*} l_j) \right] P(M_j|y).$$

Where $f_S(x|v, b, a)$ is the p.d.f. of a univariate Student-t distribution with ν degrees of freedom, location b (the mean when $\nu > 1$) and precision a (with variance $\nu / \left[a(\nu-2)\right]$

when $\nu > 2$), evaluated at x. Note $\delta = \frac{g}{g+1}$ is the shrinkage factor defined in section 2.3.1.

In addition, l_j is defined as

$$l_j = \{1 + \frac{1}{n} + \delta z'_{f,j}(Z'_j Z_j)^{-1} z_{f,j}\}^{-1},$$

$z_{f,j}$ groups the j elements of z_f corresponding to the regressors in M_j, $\hat{\beta}_j = (Z'_j Z_j)^{-1} Z'_j y$, and d^*_j is defined as

$$d^*_j = \delta y' Q_{X_j} y + \frac{1}{g+1}(y - \bar{y}\iota_n)'(y - \bar{y}\iota_n).$$

Definition 5. (Log Predictive Score) For $f = n+1, ..., n+q$, the Log Predictive Score (LPS) is defined as:

$$LPS = -\frac{1}{q} \sum_{f=n+1}^{n+q} \ln p(y_f | y). \tag{21}$$

The smaller the LPS, the better the model performs in prediction.

2.5 Role of Prior in Posterior Results

[Ley and Steel, 2009] investigate various prior structures used in the literature and examine their effects on the posterior results. In particular they measure the effects of fixing γ and letting γ be random in equation (11), choosing $m = 7$ and $m = k/2$ in equation (14), and the use of $g = n$ vs $g = k^2$ in (7).

As mentioned in section 2.3.2, they conclude that fixing γ does not constitute a good non-informative prior as it is in fact quite informative and influential on the posterior model size. [Ley and Steel, 2009] also find that the influence of the choice of m on the posterior model size is great for fixed γ, whereas its effect is much less severe for the case of random γ.

To examine the effect of a hierarchical prior on γ, we observe equation (15). This hierarchical prior structure introduces an added uncertainty on the model size. The variance in (15) multiplies the variance in the case with fixed $\gamma = m/k$ by factor

$$\frac{\frac{k}{m}+k}{\frac{k}{m}+1},$$

which for $k > 1$, is always greater than 1, and is an increasing function of m. When $m \downarrow 0$, (15) goes to 1, and when $m \uparrow k$, it goes to $(k+1)/2$.

Therefore it is clear to see this hierarchical prior structure on the model size introduces an increased variance on the model size. This helps explain why this choice has been shown to perform better as a non-informative prior.

The prior distribution on the model space only affects the posterior model inference through the prior odds (see definiton 2). For γ fixed, the prior odds are equal to

$$\frac{P(M_i)}{P(M_j)} = \left(\frac{\gamma}{1-\gamma}\right)^{k_i-k_j} = \left(\frac{m}{k-m}\right)^{k_i-k_j}, \tag{22}$$

which is equal to 1 for $\gamma = 0.5$. For $\gamma < 0.5$, the prior odds induces a prior penalty on the larger model and for $\gamma > 0.5$, a prior penalty is induced on the smaller model. The second equality is obtained by substituting $\gamma = m/k$. Equivalently, $m < k/2$ favours the smaller model and vice versa.

With the prior on M_j defined as in equation (12), and taking $a = 1$ and $b = (k-m)/m$ in equation (13), as discussed in section 2.3.2, we can obtain a form of the prior odds for random γ as such:

$$\frac{P(M_i)}{P(M_j)} = \frac{\Gamma(1+k_i)}{\Gamma(1+k_j)}\frac{\Gamma(\frac{k-m}{m}+k-k_i)}{\Gamma(\frac{k-m}{m}+k-k_j)}. \tag{23}$$

Plotting the prior odds defined in (22) and (23), and taking logarithms, similar to Figure 2 in [Ley and Steel, 2009], helps illustrate the difference between fixed and random γ. In this case, in Figure 3 we take the number of total covariates to be $k = 50$, and we take the number of covariates in the model in the numerator of the prior odds to be $k_i = 10$.

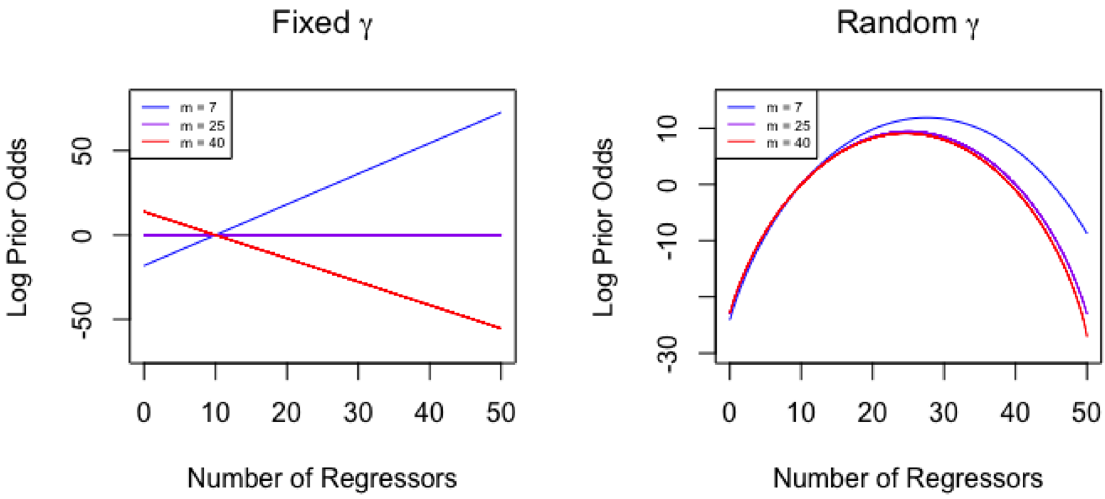


Figure 3: Log Prior Odds for $k_i = 10$ vs varying k_j

The higher the log odds, the more model M_i is favoured of model M_j.

Random γ leads to the downweighing of models around the $k_j = k/2 = 25$ point for all values of m. This counteracts the fact that most models appear in this area of the model space $\mathcal{M}$ - as $\binom{k}{k/2}$ is the largest binomial coefficient. For fixed γ, larger models are favoured for $m > k/2$, smaller models are favoured for $m < k/2$, whereas $m = k/2$ means indifference on model size. Note also the range of log prior odds taken by fixed and random γ, implying a much more informative prior stance for fixed γ as model sizes away from m are punished more strongly than in the random case. Thus, the choice of m is much more important in the fixed case than in the random case. Random γ prior adapts to the data much more naturally.

To examine the effect of g, we look at the Bayes' factor between two models M_i and M_j, defined in Definition 1 and given by the marginal likelihood structure in (10). Provided g does not depend on k_j, the Bayes' factor is

$$\frac{l_y(M_i)}{l_y(M_j)} = (g+1)^{\frac{k_j - k_i}{2}} \left(\frac{1 + g(1 - R_i^2)}{1 + g(1 - R_j^2)} \right)^{-\frac{n-1}{2}}, \tag{24}$$

where R_i^2 is the "R-squared" for model M_i defined as $R_i^2 = 1 - [y' Q_{X_i} y / (y - \bar{y}\iota_n)'(y - \bar{y}\iota_n)]$.

Equation (24) can be seen as the relative weight the data assigns to the corresponding models. This expression clearly depends on n, g, k_i, k_j and R_i^2. The Bayes' factor is crucial in understanding the differences between choices of g.

If we assume both models M_i and M_j fit equally well (i.e. $R_i^2 = R_j^2$), the Bayes' factor becomes:

$$\frac{l_y(M_i)}{l_y(M_j)} = (g+1)^{\frac{k_j - k_i}{2}}.$$

However, g tends to be relatively large (remember we are comparing $g = n$ and $g = k^2$),

$$\implies \frac{l_y(M_i)}{l_y(M_j)} \approx g^{\frac{k_j - k_i}{2}}.$$

Thus, if one of the models contains one more regressor, the penalty induced on it will be approximately equal to $1/g^{1/2}$, i.e., its Bayes factor is scaled by that amount. The penalty on the Bayes' factor would be approximately $1/g$ if it has two more regressors. It is therefore clear to see that $g = k^2$ induces a larger penalty on model size than $g = n$. This will naturally lead to the posterior results having smaller models for a prior with $g = k^2$. For a dataset with a large number of covariates, the choice of $g = k^2$ can be quite handy.

In addition, [Ley and Steel, 2009] have concluded that $g = n$ may cause convergence problems in the MCMC chain where there is relatively large k in comparison to n (as not enough of a model size penalty is imposed). They recommend the use of $g = k^2$ in such contexts.

For the hyperprior on g in [Ley and Steel, 2012], they argue that this hyperprior can have a large effect on the induced penalties for model complexity, but does not affect the relative fit of the models. This prior structure is more suitable where a large model complexity penalty is not preferred.

2.6 Jointness in the Bayesian Framework

Jointness can be thought of as the tendency for variables to appear together (complements), and disjointness can be thought of the tendency of variables to not appear together (substitutes). In essence, variables that are disjoint convey similar information and a model should not include the disjoint combination.

A set of criteria was proposed by [Ley and Steel, 2007] that a useful measure of jointness should have. The four criteria proposed are such as:

1. **Interpretability** - a jointness measure should have a formal statistical or a clear intuitive meaning of jointness.

2. **Calibration** - the jointness values should be calibrated against a clearly defined scale, derived from either formal statistical or intuitive arguments.

3. **Extreme Jointness** - where extreme jointness is present, the jointness measure should reach its value reflecting maximum jointness.

4. **Definition** - the jointness measure should always be defined whenever at least one of the variables considered is included with positive probability.

[Ley and Steel, 2007] test various different measures of jointness in the Bayesian framework against the set of criteria proposed, and conclude that the measures:

$$\mathcal{J}^*_{ij} = \frac{P(i \cap j)}{P(i \cup j)} = \frac{P(i \cap j)}{P(i) + P(j) - P(i \cap j)} \in [0, 1], \tag{25}$$

$$\mathcal{J}_{ij} = \frac{P(i \cap j)}{P(i \cap j^c) + P(j \cap i^c)} = \frac{P(i \cap j)}{P(i) + P(j) - 2P(i \cap j)} \in [0, \infty), \tag{26}$$

are suitable.

$\mathcal{J}^*_{ij}$ can be thought of as the joint probability relative to the probability of including either one. Whereas, $\mathcal{J}_{ij}$ can be thought of as the joint probability relative to the probability of including either one, but not both.

[Ley and Steel, 2007] also mention that $\mathcal{J}_{ij}$ is their preferred measure of jointness as $\mathcal{J}_{ij}$ corresponds to the posterior odds ratio of the models including both i and j vs the models that include them only individually, leading to a more straightforward statistical interpretation.

In the case of more than two regressors, the measure used by [Ley and Steel, 2007] can easily be extended. It is defined as such:

$$\mathcal{J}^*_{\mathcal{S}} = \frac{P(\mathcal{S})}{P(\subset \mathcal{S}) + P(\mathcal{S})} \in [0, 1], \tag{27}$$

$$\mathcal{J}_{\mathcal{S}} = \frac{P(\mathcal{S})}{P(\subset \mathcal{S})} \in [0, \infty), \tag{28}$$

defined by [Ley and Steel, 2007] as the *multivariate jointness* measure, where $\mathcal{S}$ is the set of regressors, $P(\mathcal{S})$ the total posterior probability assigned to models having all regressors in $\mathcal{S}$, and $P(\subset \mathcal{S})$ the total posterior probability assigned to all models including only proper subsets of $\mathcal{S}$.

Note it should be understood that all posterior quantities are conditioned on the observed sample y.

3 MCMC Methods

3.1 Introduction to MCMC Methods

In economic contexts, large model spaces are often the norm. In fact, too large for even the most advanced CPUs to handle. Consider the case where we are modelling the growth of a select number of countries that we have data for, say $n = 100$, and 50 potential covariates we are testing, so $k = 50$. It is not hard to check that in this case, the model space $\mathcal{M}$ (all potential models) has $|\mathcal{M}| = 2^k = 2^{50} = 1.126 \times 10^{15}$ or 1,125,899,906,842,624 models!

To tackle this problem, modern Markov Chain Monte Carlo (MCMC) methods are used. In the 1950s MCMC methods were invented and revolutionised fields such as statistical computing, physics and experimental mathematics. Although Monte Carlo methods are a recent development, they still predate the digital computer. Historically, the main drawback of Monte Carlo methods was that they were expensive to carry out, however this changed fundamentally with the advent of the digital computer [Johansen, 2020]. The name "Monte Carlo" goes back to Stanislaw Ulam, who claimed to be fascinated by poker and whose uncle once borrowed money from him to go gambling in Monte Carlo [Ulam, 1991]. Monte Carlo methods were initially primarily used and analysed by physicists; it was not until the 1980's, following the proposal of the Gibbs sampler in [Geman and Geman, 1984], that Monte Carlo methods featured strongly in statistics [Johansen, 2020].

Methods which generate a Markov chain, whose stationary distribution is the distribution of interest, are called Markov Chain Monte Carlo methods.

MCMC methods have increased the popularity of Bayesian statistics, as MCMC methods could now construct random sampling algorithms from a probability distribution that would enable the calculation of various Bayesian models [Hackenberger, 2019]. MCMC methods are used to map distributions that are too difficult to compute directly. These methods try to generate samples in such a way that the importance weights $\{\tilde{w_1}, ..., \tilde{w_n}\}$ associated with each sample are constant.

According to [Speagle, 2019], MCMC accomplishes this by creating a chain of (correlated) parameter values $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ over n iterations such that the number of iterations $m(\boldsymbol{X}_i)$ spent in any particular region $\delta_{\boldsymbol{X}_i}$, centred on $\boldsymbol{X}_i$, is proportional to the posterior density $p_{\boldsymbol{X}|\boldsymbol{y}}(X|y)$ contained within that region. In other words, the MCMC sampling density

$$\rho(\boldsymbol{X}) \equiv \frac{m(\boldsymbol{X})}{n},$$

at position $\boldsymbol{X}$ integrated over $\delta_{\boldsymbol{X}}$ is approximately

$$\int_{\boldsymbol{X}\in\delta_{\boldsymbol{X}}} p_{\boldsymbol{X}|\boldsymbol{y}}(\boldsymbol{X}|y)d\boldsymbol{X} \approx \int_{\boldsymbol{X}\in\delta_{\boldsymbol{X}}} \rho(\boldsymbol{X})d\boldsymbol{X} \approx \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[\boldsymbol{X}_j \in \delta_{\boldsymbol{X}}],$$

where $\mathbb{1}$ is the indicator function.

According to the Law of Large Numbers, $\rho(\boldsymbol{X}) \to p_{\boldsymbol{X}|\boldsymbol{y}}(\boldsymbol{X}|y)$ as $n \to \infty$.

Often in MCMC methods, burn-ins are used for the algorithms. This is where the first few samples (usually a predetermined proportion of the iterations) are discarded – traditionally done to mitigate for the cases where the chain starts at a "bad" starting point and oversamples low-probability regions.

There are various MCMC algorithms used in statistical computing, ranging from the Metropolis-Hastings algorithm to the Gibbs sampler to Hamiltonian Monte Carlo.

Algorithm 1. (Metropolis-Hastings) [Johansen, 2020]

Starting with $\mathbf{X}^{(0)} := (X_1^{(0)}, ..., X_p^{(0)})$ iterate for $t = 1, 2, ...$

1. Draw $\mathbf{X} \sim q(\cdot|\mathbf{X}^{(t-1)})$.
2. Compute

$$\alpha(\mathbf{X}|\mathbf{X}^{(t-1)}) = \min\left\{1, \frac{f(\mathbf{X})q(\mathbf{X}^{(t-1)}|\mathbf{X})}{f(\mathbf{X}^{(t-1)})q(\mathbf{X}|\mathbf{X}^{(t-1)})}\right\}.$$

3. With probability $\alpha(\mathbf{X}|\mathbf{X}^{(t-1)})$, set $\mathbf{X}^{(t)} = \mathbf{X}$, otherwise set $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$.

Note p indicates the dimension of $\mathbf{X}$, and $f(\mathbf{X})$ is the target distribution at $\mathbf{X}$. If the algorithm rejects the newly proposed sample $\mathbf{X}$, it stays at its current value $\mathbf{X}^{(t-1)}$. The probability the Metropolis-Hastings algorithm accepts state $\mathbf{X}$, given that it is currently in state $\mathbf{X}^{(t-1)}$ is:

$$a(\mathbf{x}^{(t-1)}) = \int \alpha(\mathbf{x}|\mathbf{x}^{(t-1)})q(\mathbf{x}|\mathbf{x}^{(t-1)})d\mathbf{x}.$$

A special case of the Metropolis-Hastings algorithm is the Random-Walk Metropolis, which often uses a normal proposal distribution

$$q(\cdot|\mathbf{X}^{(t-1)}) = \mathcal{N}(\cdot|\mathbf{X}^{(t-1)}, \sigma_q^2),$$

where σ_q^2 is some proposal variance which should not be too small nor too large.

The Metropolis-Hastings algorithm generates a Markov chain. However, as the dimension of $\mathbf{X}$ grows, the performance of Metropolis-Hastings becomes poor [Everitt, 2022].

Algorithm 2. (Gibbs Sampler) [Johansen, 2020]

```
Starting with $\mathbf{X}^{(0)} := (X_1^{(0)}, ..., X_p^{(0)})$ iterate for $t = 1, 2, ...$
1. Draw $X_1^{(t)} \sim f_{X_1|\mathbf{X}_{-1}}(\cdot|X_2^{(t-1)}, ..., X_p^{(t-1)})$.
...
$j$. Draw $X_j^{(t)} \sim f_{X_j|\mathbf{X}_{-j}}(\cdot|X_1^{(t-1)}, ..., X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)}, ..., X_p^{(t-1)})$.
...
$p$. Draw $X_p^{(t)} \sim f_{X_p|\mathbf{X}_{-p}}(\cdot|X_1^{(t-1)}, ..., X_{p-1}^{(t-1)})$.
```

Note $\mathbf{X}_{-j}$ is the vector of $\mathbf{X}$ without the j^{th} entry, called the full conditional. The Gibbs sampler is used when the dimension p of $\mathbf{X}$ is large. Where Metropolis-Hastings fails for large p, the Gibbs sampler is used instead. However, when the $\mathbf{X}$ variables are strongly dependent, the Gibbs sampler explores the space very slowly [Everitt, 2022].

3.2 MCMC Methods for Bayesian Model Averaging

In BMA, MCMC algorithms seek to generate samples proportional to the posterior model probabilities to arrive at an optimal estimate for the weights, and thus the model averaged quantities. A popular strategy is to run an MCMC algorithm in the model space, sampling the models that are most promising. Note this is performed over the discrete model space $\mathcal{M}$.

The most commonly used strategy is a random-walk Metropolis sampler, usually referred to as MC^3 [Steel, 2020]. MC^3, introduced in [Madigan et al., 1995], uses a Metropolis sampler which proposes models from a small neighbourhood around the current model M_j, namely all models with one covariate less or more [Steel, 2020].

Algorithm 3. (MC^3 Algorithm)

Let $X^{(t)}$ be the current state of the chain at time t, where X can represent any M_j for $j = 1, ..., J$, where J is the number of models in model space $\mathcal{M}$.

1. Choose coordinate i of $X^{(t)}$ using the uniform selection probabilities d and propose the new model $X = (X_1^{(t-1)}, ..., 1 - X_i^{(t-1)}, ..., X_p^{(t-1)})$.

2. Accept model X with probability:

$$\alpha(X, X^{(t-1)}) = \min \left\{ 1, \frac{l_y(X)}{l_y(X^{(t-1)})} \right\}.$$

Note $l_y(X)$ represents the marginal likelihood for model X, and X_i the indicator for covariate i in model X. Note Algorithm 3 is taken largely from Algorithm 1 in [Lamnisos et al., 2013], but modified to fit the notation in this document.

In the implementation of MC^3, the chain is used to identify the models with high posterior probability, denoted by “Bayesian Random Search” in [Lee, 1996]. [Fernandez et al., 2001b] propose to use this idea of “Bayesian Random Search” as a diagnostic aid for assessing the performance of the chain. “A high positive correlation between posterior model probabilities based on the empirical frequencies of visits in the chain on the one hand, and the exact marginal likelihoods on the other hand, suggests that the chain has reached its equilibrium distribution” [Fernandez et al., 2001b]. In the case of an extremely large model space, it is unlikely the chain will reach all models, however the objective of sampling methods in this context is to explore the model space to identify the models with higher posterior probability.

When marginal likelihoods (and parameter posteriors) are not available analytically, an MCMC chain can be used to sample from the parameter space alongside the model space. In particular, the Gibbs sampler defined in Algorithm 2 can be used to sample from the joint space of the parameters and the models, as used in [George and McCulloch, 1993]. This procedure is often denoted as Stochastic Search Variable Selection (SVSS) [Steel, 2020]. Whenever the marginal likelihood is available analytically (such as for the normal linear likelihood, prior structure 1 on the parameters and fixed g), it suffices to use an MCMC algorithm, such as MC^3, to sample over just the model space.

4 Other Model Averaging Methods

4.1 Frequentist Model Averaging

In frequentist statistics, the parameter is a fixed quantity, as opposed to being a random variable in Bayesian statistics. In addition, frequentist methods tend to focus on parameters and their properties. They do not assign a prior, as Bayesian methods do. Frequentist methods rely on repeated sampling to obtain estimates of parameters. Bayesian methods assign prior knowledge on parameters and use the data to obtain probability distributions on them. Hence, the availability of marginal and predictive distributions for Bayesian methods.

We can describe Frequentist Model Averaging (FMA) as

$$\hat{\beta}_{\text{FMA}} = \sum_{j=1}^{J} \omega_j \hat{\beta}_j, \tag{29}$$

where $\omega_j \in [0, 1]$ $\forall j = 1, ..., J$, are model weights, and $\hat{\beta}_j$ is an estimator based on model M_j.

The question then becomes how to decide on these weights. Whereas in BMA, the weights can be naturally defined as the posterior model probabilities, in FMA there is no such obvious approach. As such, there have been many proposals for such weights in the literature.

[Hansen, 2007] propose selecting model weights by minimising Mallow's C_p criterion. This model averaging method is called Mallow's Model Averaging (MMA) and has various asymptotic optimality properties, as the Mallows' C_p criterion is asymptotically equivalent to the squared predictive error [Steel, 2020]. [Hansen, 2007] assumes a standard linear regression model, and estimates the l^{th} model $y = Z_l \beta_l + \epsilon_l$, by least squares. This is done for all models $M_j, j = 1, ..., J$, and the final estimator of β is

$$\hat{\beta}_{\text{MMA}} = \sum_{j=1}^{J} \omega_j \hat{\beta}_j,$$

where $\omega_j, j = 1, ..., J$ are normalised weights, calculated by minimizing the Mallow's criterion:

$$C_p(\Omega) = \Omega'\hat{\epsilon}'\hat{\epsilon}\Omega + 2\sigma^2 L'\Omega,$$

with $\Omega_{J\times 1} = (\omega_1, ..., \omega_J)$ representing the vector of weights, $\hat{\epsilon}_{n\times J} = (\hat{\epsilon}_1, ..., \hat{\epsilon}_J)$ the matrix of all residual vectors, and $L_{J\times 1} = (k_1, ..., k_J)'$ the vector of the number of regressors used for each of the J models. The optimisation problem is as such: [Amini and Parmeter, 2012]

$$\min_{\omega_j} C_p(\Omega) = \Omega'\hat{\epsilon}'\hat{\epsilon}\Omega + 2\sigma^2 L'\Omega,$$

$$\text{s.t.} \sum_{j=1}^{J} \omega_j = 1, \;\; \omega_j \in [0, 1] \; \forall j = 1, ..., J,$$

where σ^2 can be replaced with the estimator $\hat{\sigma}_J^2 = (n - J)^{-1}\hat{\epsilon}_J'\hat{\epsilon}_J$.

[Hansen and Racine, 2012] propose selecting model weights through minimising a leave-one-out cross-validation criterion. This model averaging method is called Jackknife Model Averaging (JMA) and like MMA, has asymptotic optimality properties. In particular, reaching the lowest possible squared errors over the class of linear estimators. However, unlike MMA, JMA has optimality properties even under heteroscedastic errors, and when the candidate models are non-nested [Steel, 2020].

[Buckland et al., 1997] propose defining the weights as such:

$$\omega_j = \frac{\exp(-I_j/2)}{\sum_{j=1}^{K} \exp(-I_j/2)}$$

Where I_j is an information criterion, such as the AIC or BIC criteria.

Unlike most FMA asymptotically optimal methods that consider normal linear models, [Zhang et al., 2016] consider Generalised Linear Models and propose weights based on a plug-in estimator of the Kullback-Leibler loss plus a penalty term [Steel, 2020]. They prove asymptotic optimality for fixed and growing number of covariates.

As FMA does not lead to model probabilities, simple MCMC algorithms that visit models in line with their probabilities are not readily available, unlike with BMA. This means that FMA can often not deal with large model spaces, whereas we have shown BMA can. Various researchers of FMA have proposed many different ways to attempt to reduce the model space to allow for FMA methods to take place. Common methods include applying orthogonal transformations of the regressors, always including subsets of covariates and conducting a preliminary model-screening step to remove the least interesting models [Steel, 2020].

4.2 Mixture of Bayesian and Frequentist Model Averaging

[Magnus et al., 2010] proposes a Weighted Average Least-Squares (WALS) estimator. This can be seen as a mixture of BMA and FMA procedures. In essence, the weights it assigns in equation (29) are given a Bayesian justification, however inference on the model space is not made in the Bayesian framework, but rather in the frequentist framework. The weights are based on risk considerations and can be determined in a Bayesian or frequentist sense. [Magnus and De Luca, 2016] state: "The WALS procedure surveyed in this paper is a Bayesian combination of frequentist estimators. The parameters of each model are estimated by constrained least squares, hence frequentist. However, after implementing a semiorthogonal transformation to the auxiliary regressors, the weighting scheme is developed on the basis of a Bayesian approach in order to obtain desirable theoretical properties such as admissibility and a proper treatment of ignorance". Thus, we can view WALS as a frequentist model builder, with Bayesian averaging. They also state that WALS cannot deal with variable jointness, as WALS does not provide posterior inclusion probabilities, which the available measures of jointness are based on (see section 2.6). However, it seems that WALS can give similar results to BMA. [Amini and Parmeter, 2012] compare BMA, WALS and MMA on datasets from; [Fernandez et al., 2001b], [Masanjala and Papageorgiou, 2008] and [Doppelhofer and Weeks, 2009], and conclude that that all three approaches give similar results. They state: "We found that the sign and magnitude of both the WALS and the MMA coefficients were roughly identical to those obtained via BMA. We are encouraged by the similarity of these three methods". However, they also state: "Traditional BMA analyses also focus attention on the PIPs [posterior inclusion probabilities] as well as measuring jointness between pairs of covariates. In the current setup, neither WALS nor MMA provides this information. Future work to derive similar metrics from WALS and MMA are required before we have the ability to favor one method over another categorically".

[Sala-i Martin et al., 2004] proposes a Bayesian Averaging of Classical Estimates

(BACE) model averaging technique. They show that the weighting method can be derived as a limiting case of a standard Bayesian analysis as the prior becomes "dominated" by the data. They state: "BACE combines the averaging of estimates across models, which is a Bayesian concept, with Classical OLS estimation which comes from the assumption of diffuse priors". "Classical estimates" refers to model building methods most econometricians are familiar with, and "Bayesian Averaging" refers to the Bayesian approach of model averaging taken. BACE requires the specification of only one prior hyper-parameter, the mean model size k, and relies on an approximation as sample size n goes to infinity (strong assumption for growth context), to arrive at a BIC approximation. The weights given to different models are proportional to the log of the likelihood function corrected for degrees of freedom. According to [Ley and Steel, 2009] the Bayes' factor given in [Sala-i Martin et al., 2004] is as such:

$$\frac{l_y(M_i)}{l_y(M_j)} = n^{\frac{k_j-k_i}{2}} \left(\frac{1-R_i^2}{1-R_j^2} \right)^{-\frac{n}{2}},$$

which is similar to the Bayes' factor in equation (24), if we take $g = n$:

$$\frac{l_y(M_i)}{l_y(M_j)} = (n+1)^{\frac{k_j-k_i}{2}} \left(\frac{1+\frac{1}{n}-R_i^2}{1+\frac{1}{n}-R_j^2} \right)^{-\frac{n-1}{2}}.$$

5 Real-data Application

5.1 Posterior Results

We use data from [Bruns and Ioannidis, 2020], for $n = 63$ countries and $k = 37$ possible explanatory variables. The outcome variable is the annualised average growth rate of GDP per capita for the 20-year growth period between 1990 and 2010. In this dataset we focus on modelling the GDP rate of growth, which will inevitably lead to different assumptions than modelling the GDP levels according to prevailing economic theory.

We shall consider all combinations of regressors; thus, we have a model space size J of $2^k = 2^{37} = 137{,}438{,}953{,}472$ models. [Bruns and Ioannidis, 2020] provide data for 20 and 35-year growth periods between 1960 and 2010 and they compare model outputs for different periods. However, in this project we only consider the most recent growth period available to better illustrate BMA in practice. Table 1 provides a description of the regressors considered. We consider the normal linear regression model defined in equation (6) and the prior structure on the parameters defined in (7). Further, we want to induce a strong penalty on model complexity (as we have large k) thus we take g to be fixed and equal to the benchmark prior defined in [Fernandez et al., 2001a]. We have $g = k^2$ as $k^2 \gg n$ (see section 2.3.1). In addition, we do not have a strong prior stance on the expected model size, thus we take a hyper-prior on the inclusion probability γ, as in equation (11), and take the prior mean model size m to be $m = 7$.

We use [Zeugner and Feldkircher, 2015]'s BMS package in R. For our MCMC sampler, we use the MC^3 sampler (see Algorithm 3) that wanders through the model space by adding or dropping regressors from the current model. We perform 2 million iterations and 500,000 burn-ins, and we store the top 2000 models.

Table 2 provides the posterior inclusion probabilities (PIPs) of each variable. As is evident, the variables that are included in the top models are Population (with almost a probability of 1), Log GDP Per Capita, Life Expectancy and Sub-Saharan Africa. Additionally, other variables such as Christianity, Fertility, Islam, Average Growth Rate of Terms of Trade, Investment Price, Total GDP, Polity, Europe, and Population Density have non-negligible PIP.

Table 1: Description of Data

Outcome Variable	
Growth Rate of GDP Per Capita:	Average annual growth rate of real GDP per capita for period 1990-2010 at constant 2005 national prices (2005US$).
Geographic Region	
Sub-Saharan Africa:	Indicator variable for Sub-Saharan African countries.
Europe:	Indicator variable for European countries.
East Asia:	Indicator variable for East-Asian countries.
Latin America:	Indicator variable for Latin American countries.
Climate	
Latitude:	Absolute Latitude
Tropical Climate Zone:	Fraction tropical climate zone.
Fraction Tropics:	Fraction tropical climate zone.
Colonial History	
British Colony:	Indicator variable for former British colonies after 1776.
Spanish Colony:	Indicator variable for former Spanish colonies.
Colony:	Indicator variable for former colonies.
Education	
Human Capital:	Index of human capital per person based on years of schooling and returns to education.
Primary Schooling:	Share of population over 15 that is enrolled in or finished primary school.
Tertiary Education:	Share of population over 15 that completed tertiary schooling.
Demography	
Population:	Population in millions.
Population Density:	Population density (people per sq. km of land area).
Fertility:	Total fertility rate (total births per woman).
Life Expectancy:	Life expectancy at birth in years.
Institutions	
Polity:	Polity2 score.
Duration Regime Change:	Number of years since the most recent regime change.

Trade	
Landlock:	Indicator variable for landlocked countries.
Land Area:	Total land area in km^2.
Distance to Major City:	Logarithm of minimal air distance (in km^2) from New York, Tokyo or Rotterdam .
Land Near Navigable Water:	Proportion of country's land area within 100 km of ocean or ocean-navigable river.
Openness:	Share of merchandise exports at current PPPs + share of merchandise imports at current PPPs (1990-2010).
Average Growth Rate of Terms of Trade:	Average annual growth rate of the ratio of price level of exports and price level of imports, with price level of USA GDPo in 2005 equal to 1 (1990-2010).
Terms of Trade:	Ratio of price level of exports and price level of imports, with price level of USA GDPo in 2005 equal to 1 (1990-2010).
Religion	
Christianity:	Percentage of adherence to Christianity.
Islam:	Percentage of adherence to Islam.
Judaism:	Percentage of adherence to Judaism.
Hindu:	Percentage of adherence to Hinduism.
Buddhism:	Percentage of adherence to Buddhism.
Natural Resources	
Primary Exports:	Share of exports of fuel and lubricants at current PPPs.
Macroeconomy	
GDP Per Capita:	Logarithm of expenditure-side real GDP per capita at current PPPs (in mil. 2005US$).
Total GDP:	Logarithm of expenditure-side real GDP at current PPPs (in mil. 2005US$).
Investment Price:	Price level of capital formation with price level of USA GDPo in 2005 equal to 1.
Investment Share:	Share of gross capital formation at current PPPs.
Government Consumption Share:	Share of government consumption at current PPPs.

Table 2: BMA Output I

Explanatory Variable	Posterior Inclusion Probability
Population	0.9995540
Log GDP Per Capita	0.5552295
Life Expectancy	0.5299280
Sub-Saharan Africa	0.2730555
Christianity	0.0254040
Fertility	0.0249280
Islam	0.0205745
Average Growth Rate of Terms of Trade	0.0186315
Investment Price	0.0149470
Total GDP	0.0135885
Polity	0.0129985
Europe	0.0123500
Population Density	0.0117010
Duration Regime Change	0.0068345
Latitude	0.0067200
Distance to Major City	0.0065580
Human Capital	0.0059945
Primary Schooling	0.0059855
Buddhism	0.0056640
Government Consumption Share	0.0054705
Spanish Colony	0.0046995
Fraction Tropics	0.0046645
Latin America	0.0043770
Primary Exports	0.0042735
East Asia	0.0041180
Land Near Navigable Water	0.0040610
British Colony	0.0038730
Tropical Climate Zone	0.0038410
Tertiary Education	0.0038355
Hindu	0.0038085
Openness	0.0036995
Landlock	0.0035575
Colony	0.0033615
Investment Share	0.0033330
Judaism	0.0031675
Land Area	0.0026525
Terms of Trade	0.0026440

In fitting Bayesian models, we obtain probability distributions for the parameters, unlike frequentist methods which only give us point estimates alongside measures for confidence intervals. Table 3 provides us with the expected value of the parameters, in this case the regressors' coefficients, and the standard deviation.

Figure 4 gives the conditional marginal densities of the 8 regressors with highest posterior inclusion probability. In particular, for the Population variable, our estimates are quite precise, as captured by the small standard deviation of 0.0000082.

The coefficient estimates given by this model can lend to some economic theory. For example, it is found that the Population variable has a positive coefficient (with small variance), which can imply a country's population has a strong correlation with its GDP growth. The logarithm of GDP Per Capita has a negative coefficient, which can imply a sort of convergence or stabilisation of economic development, where countries with lower GDP per capita show higher growth in GDP. In fact, this hypothesis is called the "convergence" or "catch-up effect" in economics and theorises that developing countries have the potential to grow at a faster rate than developed countries, because diminishing returns are not as strong as in the capital-rich countries.

Longer life expectancy, higher Muslim population and higher average growth rate of terms of trade are also variables linked with a higher GDP growth according to the model. Whereas Sub-Saharan African countries, higher Christian population and higher fertility are linked with a lower GDP growth.

Interestingly, variables such as Distance to Major City and Land Area are judged to have negligible effect on the GDP growth rate, represented by expected posterior means of practically zero. Moreover, the variance of their marginal densities are very small - indicating the model is "sure" of this conclusion.

Table 3: BMA Output II

Explanatory Variable	$E[\beta_i \vert y]$	$\sqrt{\text{Var}(\beta_i \vert y)}$
Population	0.0000467	0.0000082
Log GDP Per Capita	-0.0061564	0.0058457
Life Expectancy	0.0008568	0.0008402
Sub-Saharan Africa	-0.0038243	0.0066282
Christianity	-0.0002027	0.0014087
Fertility	-0.0001292	0.0009294
Islam	0.0001820	0.0014521
Average Growth Rate of Terms of Trade	0.0044359	0.0360803
Investment Price	-0.0000977	0.0008983
Total GDP	-0.0000228	0.0002597
Polity	-0.0000072	0.0000741
Europe	-0.0000908	0.0009552
Population Density	-0.0000003	0.0000031
Duration Regime Change	-0.0000004	0.0000061
Latitude	0.0000006	0.0000130
Distance to Major City	0.0000000	0.0000001
Human Capital	0.0000286	0.0005555
Primary Schooling	0.0000856	0.0017330
Buddhism	0.0000498	0.0009576
Government Consumption Share	0.0001478	0.0026992
Spanish Colony	0.0000066	0.0003849
Fraction Tropics	-0.0000156	0.0003517
Latin America	0.0000018	0.0003426
Primary Exports	-0.0000681	0.0014903
East Asia	0.0000226	0.0004919
Land Near Navigable Water	0.0000180	0.0004926
British Colony	0.0000099	0.0002514
Tropical Climate Zone	0.0000068	0.0004602
Tertiary Education	-0.0000396	0.0029627
Hindu	-0.0000227	0.0008075
Openness	0.0000009	0.0002832
Landlock	-0.0000121	0.0003545
Colony	0.0000044	0.0002387
Investment Share	0.0000312	0.0016840
Judaism	0.0003638	0.0258848
Land Area	0.0000000	0.0000000
Terms of Trade	-0.0000165	0.0009354

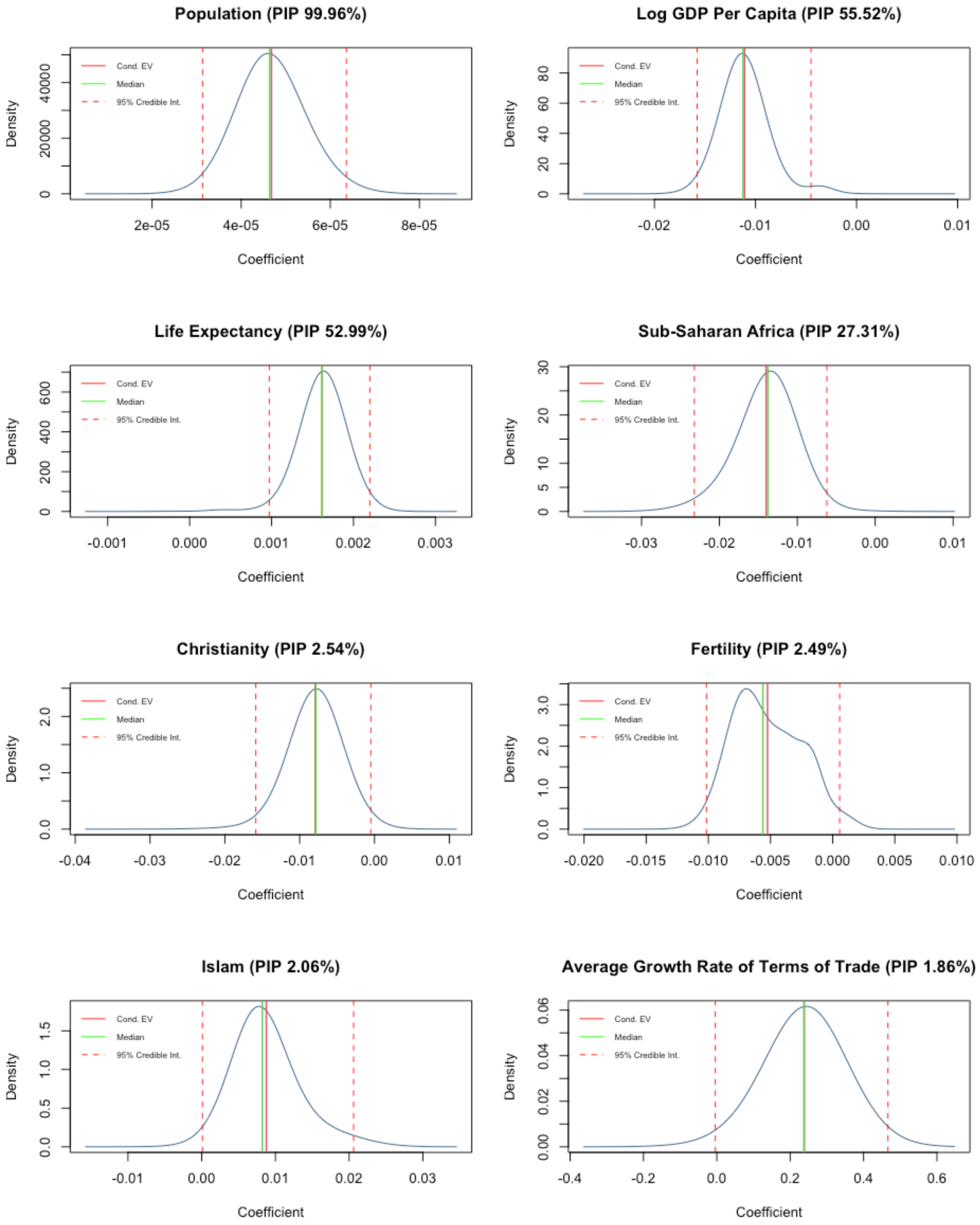


Figure 4: Marginal Densities of Key Variables

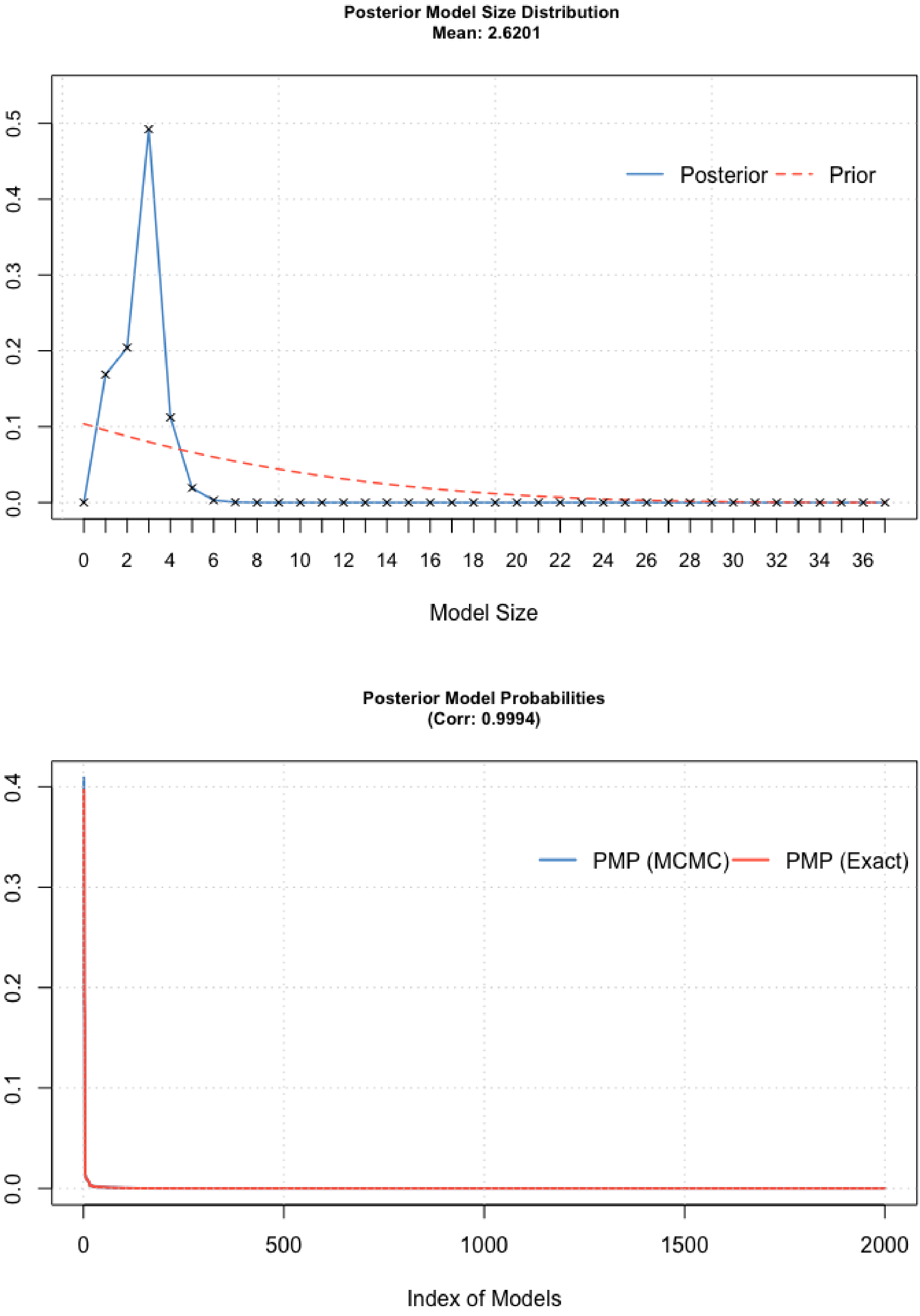
BMA addresses the issue of model probabilities, and not just of individual regressors, and thus provides a much richer type of information than just the PIPs. Figure 5 gives the posterior results on the model space. With this model prior structure, we obtain a posterior model size with mean 2.6201, despite specifying a prior model size of 7. In addition, we can see the MCMC algorithm converges to the exact posterior model probabilities (with correlation 0.9994), suggesting the algorithm performed well in identifying the models with highest probabilities.

Furthermore, our MCMC algorithm visited 64,931 models, which despite only accounting for 0.000047% of all models, corresponds to virtually all the non-negligible posterior model probabilities. The percentage of top models visited is virtually 100%, with the total posterior probabilities visited being 0.9992295. This implies our best 2000 models have 99.92295% of the total posterior probability. Despite only being included in 1137 models (roughly 1.75% of all models visited), Population is included in essentially all models with non-negligible probabilities. Further proving the convergence of the MCMC algorithm.

Table 4 gives the top 10 models and their corresponding PMPs, both exact and MCMC-estimated.

Table 4: Best Models

Best Models	PMP (Exact)	PMP (MCMC)
1	0.396184898	0.4087715
2	0.193431224	0.1841020
3	0.176124309	0.1685870
4	0.013691397	0.0136445
5	0.012871742	0.0131910
6	0.010358675	0.0101990
7	0.009474571	0.0107180
8	0.009423391	0.0086285
9	0.009019880	0.0091670
10	0.008849579	0.0087100


Posterior Model Size Distribution
Mean: 2.6201

Posterior Model Probabilities
(Corr: 0.9994)

Figure 5: Posterior Results on Model and MCMC Convergence

5.2 Predictive Results

BMA lends itself not only to inference, but also to prediction. To illustrate BMA in prediction, we randomly split the dataset into testing and training data 20 times, with the training data accounting for 75% of observations and the testing data accounting for the other 25%. We then train our BMA model on the training data for one of the data partitions, and obtain a predictive distribution (see Figure 6) for an observation in the testing data - in our case we choose Spain.

According to Figure 6, our model predicts Spain to have an expected average annual growth rate of 0.01763 (to 5 decimal places), or just under 2%, between the years 1990 and 2010. We expect the actual value to be concentrated around this estimate. The 95% credible interval for our predictive distribution is between -0.00379 and 0.03917. In reality, the actual value in the testing data is 0.01480, which is relatively close to our expected value.

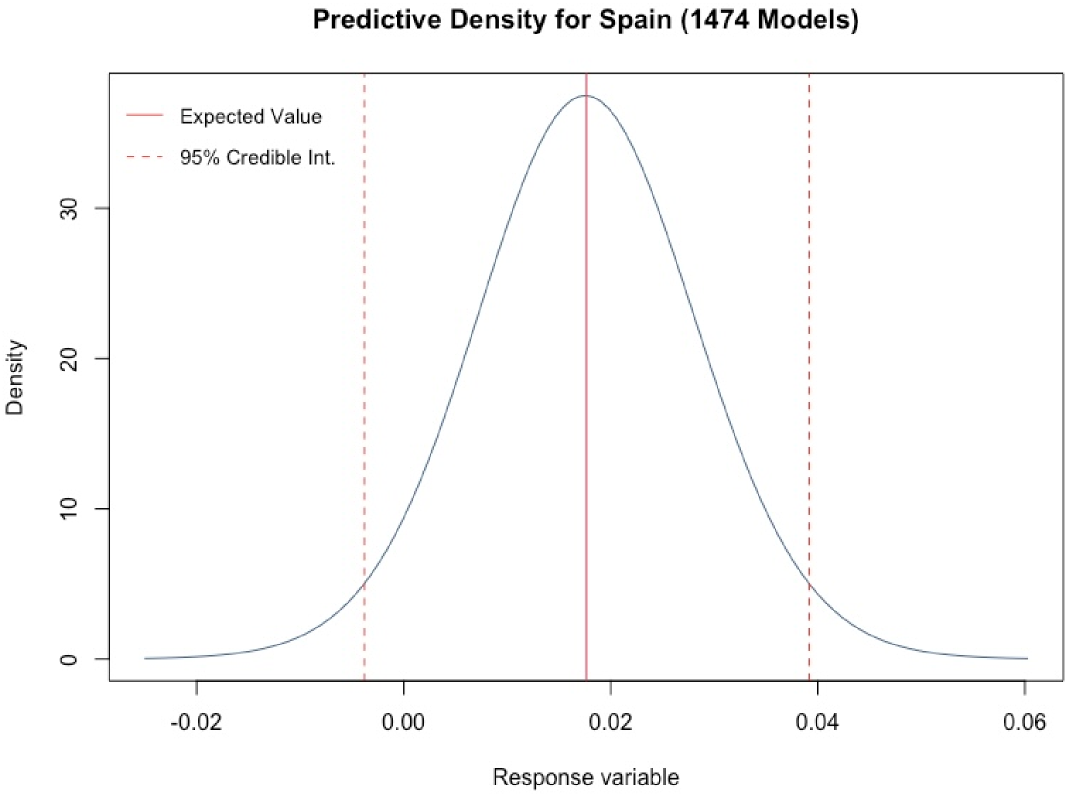


Figure 6: Predictive Distribution for Spain

5.3 Comparision of BMA to Other Methods

5.3.1 Bayesian Model Averaging vs Bayesian Model Selection

It could be argued the main question of model averaging is how it compares to model selection. In particular, how well it explains the data and how well it predicts future observations.

For inference, Bayesian Model Selection (BMS) provides a marginal distribution for the selected variables. This marginal distribution uses only information in the selected model, whereas BMA uses information from all models to form the marginal distribution.

For our best model (with highest posterior probability) defined in Table 4, we have the variables Population, Log GDP Per Capita and Life Expectancy selected. Figure 7 gives the marginal distribution of these variables conditioned on this model. In addition, the corresponding BMA marginal distributions for these variables are provided underneath each for illustration purposes.

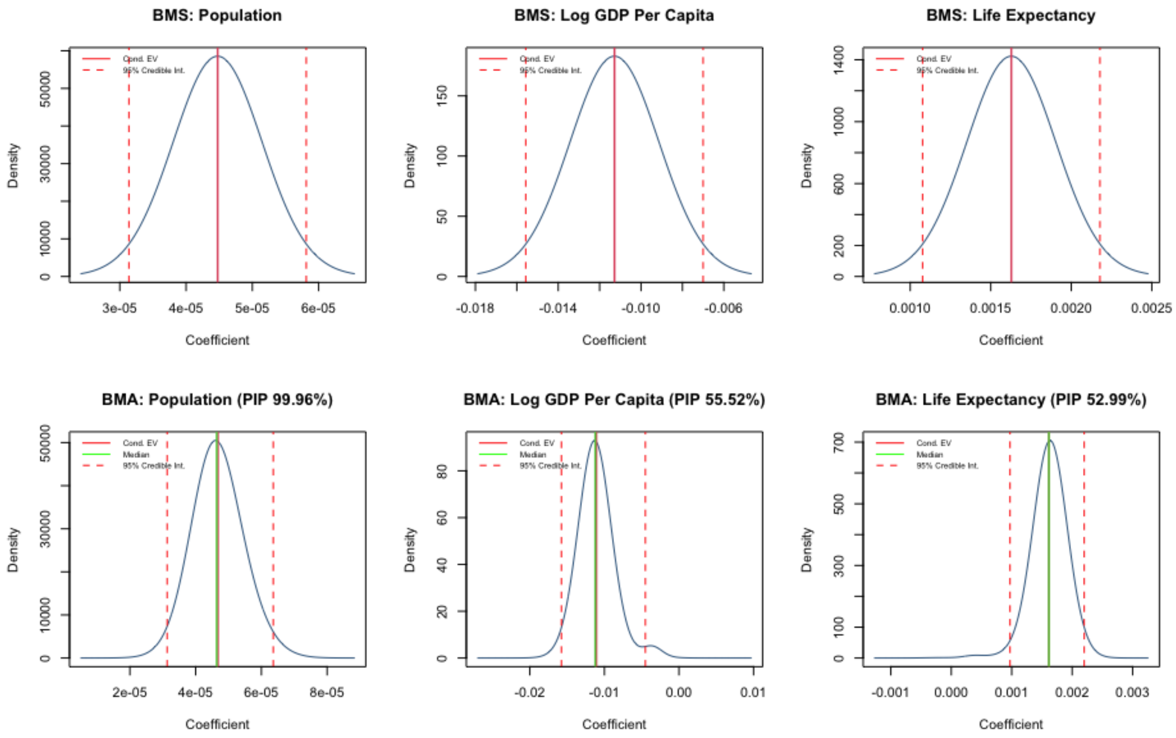


Figure 7: BMS and BMA Marginal Distributions of the Variables in the Top Model

As is evident from Figure 7, the information used from other models has managed to slightly change the marginal distributions of the three variables; however, the distributions are not changed too much. This is what we would expect considering the percentage of the weight from all remaining models is only around 60%.

For measuring the strength of prediction, we use the Log Predictive Score (defined in section 2.4) for the 20 random data partitions mentioned in section 5.2. Recall the lower the Log Predictive Score, the better the predictive distribution fits to the test observations. Table 5 gives a summary of the Log Predictive Scores for the 20 data partitions.

Table 5: BMS vs BMA For Prediction

Method	Number of Times Best Score	Min LPS	Mean LPS	Max LPS
BMS	3	-3.220056	-2.761863	-1.676565
BMA	17	-3.249651	-2.879621	-2.167746

5.3.2 Bayesian Model Averaging vs Frequentist Methods

BMA provides not only a coefficient for each variable in the dataset, but also a marginal distribution. This helps us view how the variables explain the data and how strongly they do. This feature is not provided in model selection, and in the case of frequentist methods, a marginal distribution is also not provided. The coefficient point estimate for the selected variables is instead used to determine the effect of the regressors on the outcome variable for frequentist methods.

On top of that, in comparing BMA (or any Bayesian method) to frequentist methods, we cannot compare predictive distributions - as frequentist methods do not give one. Thus, we cannot use the Log Predictive Score. Rather it is required we use a single statistic, due to frequentist methods giving point estimates. In our case we shall use the expected value of the predictive distribution as a point estimate and calculate a loss. This loss will then be compared with the loss produced from the frequentist methods. The expected value of this estimate is very similar to the classical point estimate [Zeugner and Feldkircher, 2015].

We use a single statistic for BMA, which we take to be the predictive distribution mean, and calculate the squared error loss from our predictions and actual values. We perform the same loss calculation for the frequentist methods.

We will compare Bayesian Model Averaging to stepwise variable selection, both backwards and forwards using the BIC criteria, and to LASSO regression. The squared error loss for the methods are then plotted and compared to one another. We would expect LASSO to perform well as we have a large number of covariates p and relatively few data points n. We would also expect forward stepwise selection to perform better than backwards stepwise selection, due to the tendency for backwards stepwise variable selection to overfit to the training data when there is a relatively large p to the number of data points n.

As it turns out, BMA and LASSO seem to predict the best (using the squared error loss function), followed by forward stepwise variable selection, and finally backward stepwise variable selection performs by far the poorest. This is most likely due to the large number of covariates selected in backward stepwise variable selection, leading to increased model complexity, and hence overfitting. Figure 8 gives us the squared error loss of the methods for each data partition, and Table 6 gives us the number of times each method was beaten by BMA using the squared error loss metric.

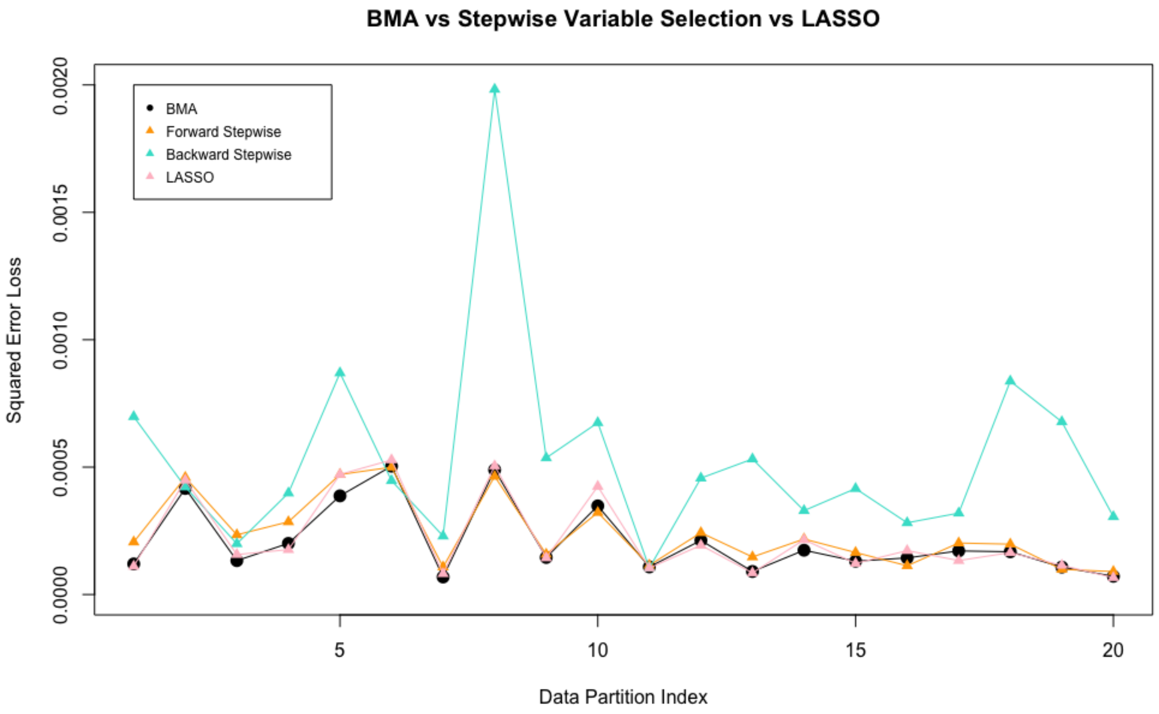


Figure 8: Squared Error Loss by Model Selection Method

Table 6: BMA vs Other Methods

Method	Number of Times Beaten by BMA
Backward Stepwise	18
Forward Stepwise	15
LASSO	10

As the data is economic growth data, there is typically not many data-points, as we are potentially limited by the number of countries (200 or so). This leads to large fluctuations in predictive accuracy when the data is split into training and testing data. However, even with this limitation, we can see BMA convincingly predicts the best alongside LASSO. Furthermore, BMA also provides us with a predictive distribution, which certainly allows us to understand model uncertainty a bit better.

5.4 Effect of Priors

As argued by [Ley and Steel, 2009], prior assumptions can have a substantial effect on the posterior results. In this section we explore a few prior settings explored in the literature, and how they affect the posterior results - as well as how they differ from the results in section 5.1.

Starting with g in equation (7), there are a few choices in the literature for this value. Recall in section 5.1, we take g to be fixed and equal to the benchmark prior defined by [Fernandez et al., 2001a] as $g = \max\{n, k^2\}$. In our dataset this corresponded to $g = k^2$, which is also known as the risk inflation criterion prior. However, the value of $g = n$, known as the unit information prior, has been proposed in the literature before. In addition to the unit information prior, a hyper-g prior has also been proposed before in the literature, as in [Liang et al., 2008].

Using the same prior setting as in section 5.1, with the exception of the values of g, we illustrate in Figure 9 the effect of g on the posterior model probabilities. Note the hyper-g prior has an expected shrinkage corresponding to the benchmark prior in [Fernandez et al., 2001a].

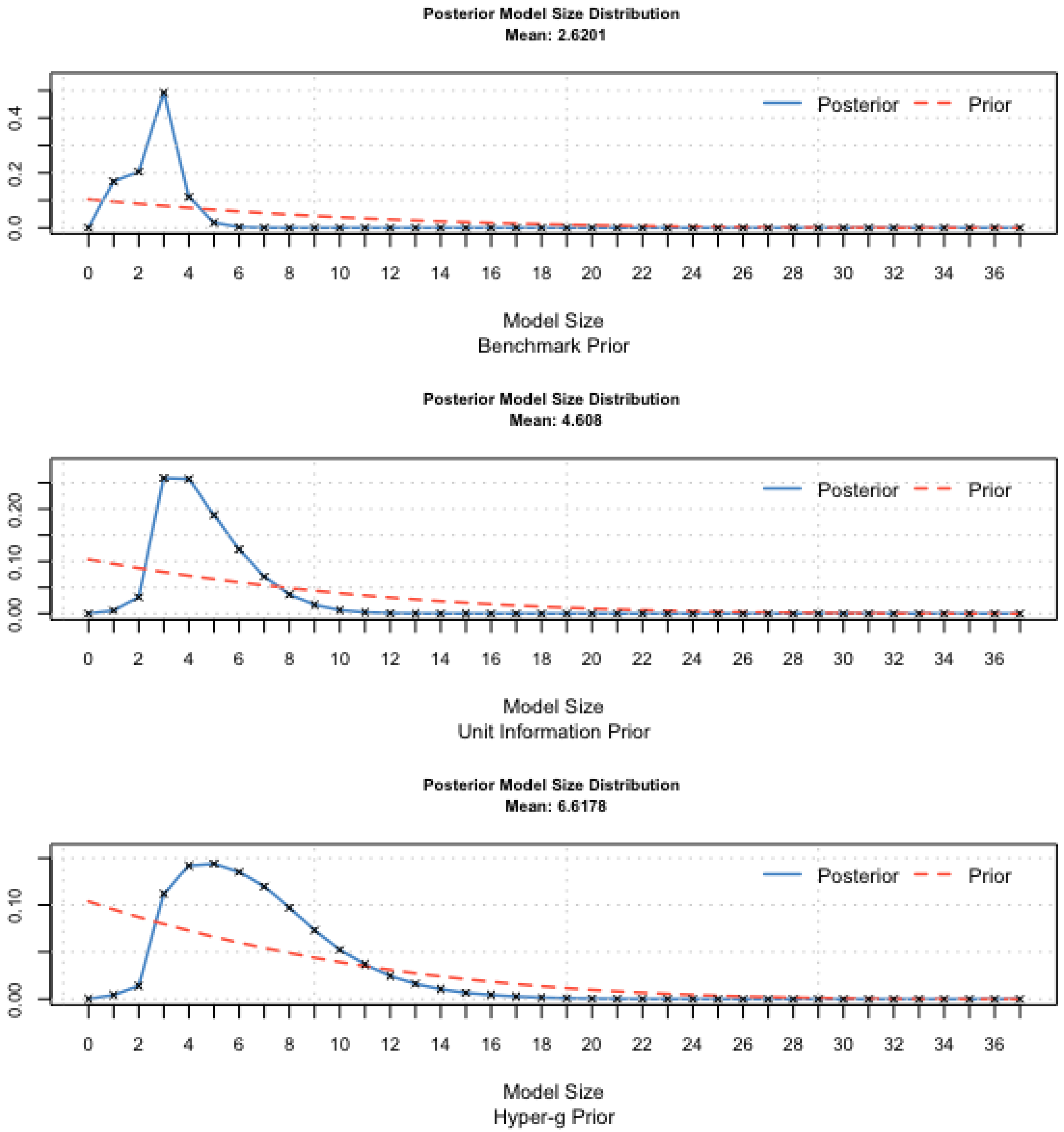


Figure 9: Effect on Posterior Model Size by Value of g

As explained in section 2.5, between a larger model and a smaller model with x regressors less, the penalty on the Bayes' factor for the larger model is approximately $1/g^{x/2}$. Thus, it is clear to see the larger the value of g, the stronger the penalty on the Bayes' factor, and the more the larger models are penalised.

Additionally, the posterior model probabilities for the unit information and hyper-g priors are more spread out, whereas for our benchmark prior, the PMP converges more to a select few models. Table 7 gives a comparison of the PMP convergence.

Table 7: Percentage of PMP of the Top 2000 Models

g	% PMP of Top 2000 Models	% PMP of Top Model (Exact)
$g = \max\{n, k^2\}$	99.9	39.6
$g = n$	81.6	20.4
Hyper-g	48.5	8.36

In addition to the effect of different g, the effect of fixing or letting γ be random and altering m is pronounced. As explained in section 2.5, fixing γ is not ideal when a non-informative prior on the model size is needed.

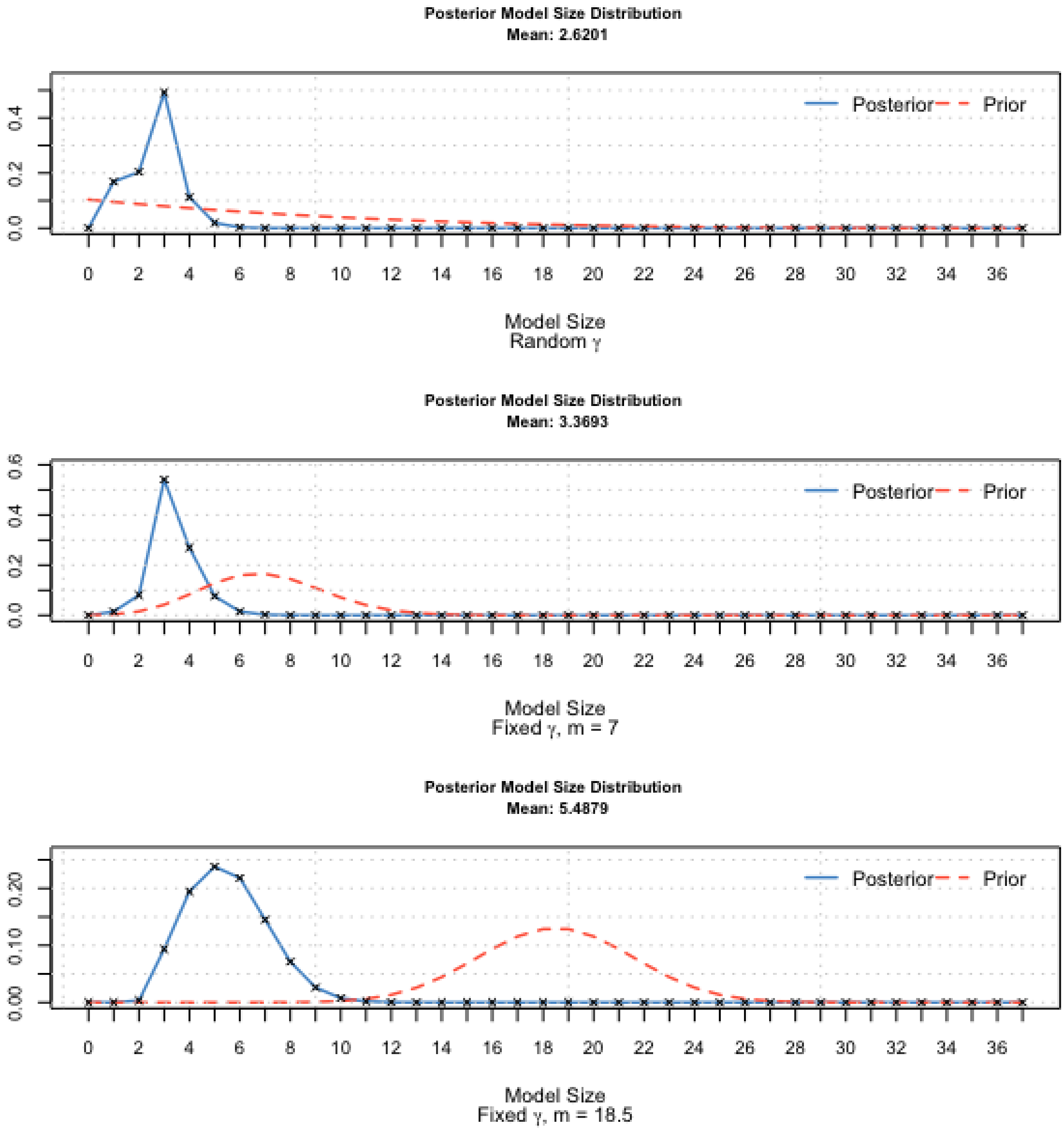


Figure 10: Effect on Posterior Model Size by Value of γ

As shown in Figure 10, fixing the value of m to $k/2 = 18.5$ has an effect of increasing the mean posterior model size. However, the posterior mean model size is still below the prior m value. This is mainly due to the shrinking effect of g.

Table 8 gives a comparison of the PMP convergence for different prior settings of γ and m.

Table 8: Percentage of PMP of the Top 2000 Models

$(\gamma,\, m)$	% PMP of Top 2000 Models	% PMP of Top Model (Exact)
γ random, $m = 7$	99.9	39.6
γ fixed, $m = 7$	98.5	44.4
γ fixed, $m = 18.5$	71.6	7.51

As the parameter values in BMA depend on $P(M_j|y)$, the averaged values of the parameters differ for different prior settings. Although the ordering of posterior inclusion probability is generally the same for the variables in the different prior settings (especially significant variables), the value of the PIP and the marginal distribution may differ slightly.

Figure 11 gives the marginal distribution of the variable with 3rd highest PIP in the different prior settings.

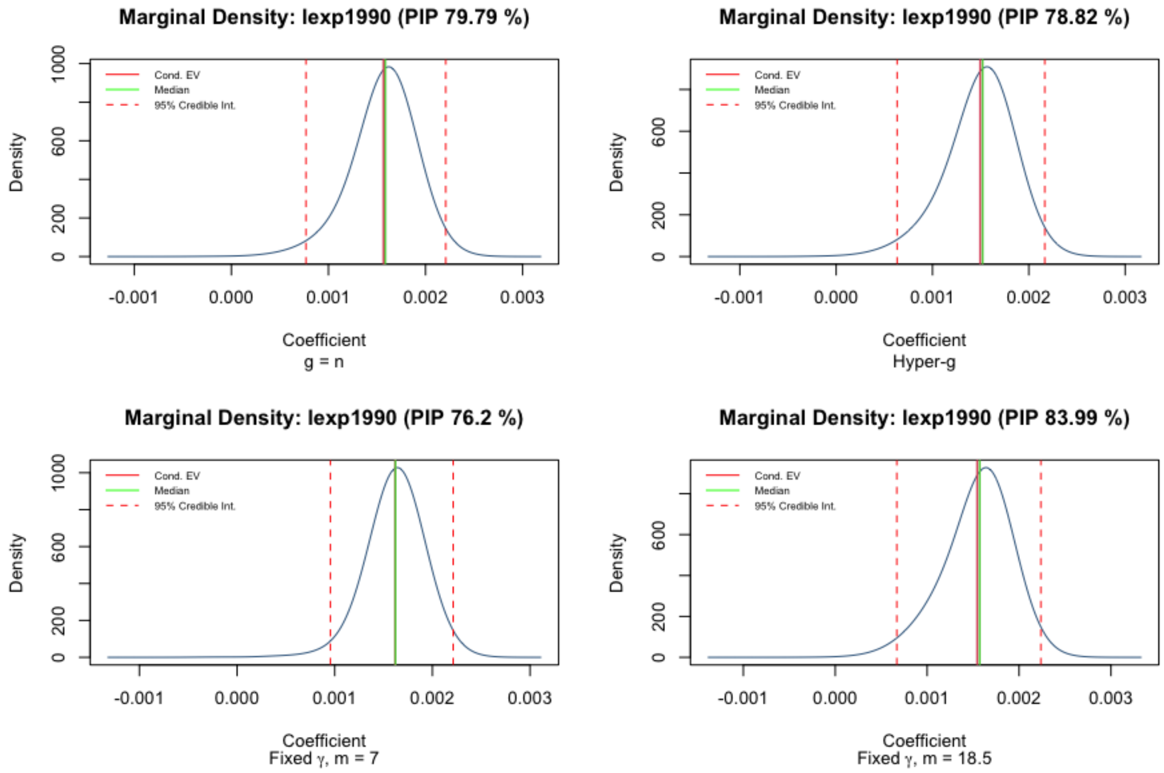


Figure 11: Marginal Distributions of Life Expectancy for Different Priors

Finally, [Ley and Steel, 2009] warn of the combination of $g = n$, fixed γ and $m = 18.5$. This combination has been shown to display weak convergence, where the MCMC algorithm struggles to adequately describe the posterior distribution on the model space $\mathcal{M}$.

To illustrate this, we use this combination on our dataset as we have above. The result is our top 2000 models only contain 6.48% of the PMP, with the top model having only 0.0558% of the PMP. This is due to $g = n$ not strongly penalising large models, fixed γ not adjusting well to the data, and $m = 18.5$ being in the area of the model space with the most possible models and covariate combinations - due to the central binomial coefficient $\binom{2n}{n}$ having the largest value.

6 Extension to Generalised Linear Models

6.1 The Model

Generalised Linear Models (GLMs) describe a more general class of models. A normal linear model has the form: [Alili, 2022]

$$Y \sim N(\mu, \sigma^2), \quad \textbf{(random part);}$$

$$\mu = Z'\beta, \qquad \textbf{(systematic part).}$$

Definition 6. (Generalised Linear Model) [Alili, 2022]
A GLM generalises both parts:

1. Y still has mean μ, but may be non-Normal
2. $g(\mu) = Z'\beta$, where $g(\cdot)$ is not necessarily the identity function.

$g(\cdot)$ is called the link function, and $Z'\beta$ is called the linear predictor.

In a GLM, the outcome y is assumed to be generated from a particular distribution from the exponential family [Li and Clyde, 2018].

Definition 7. (Exponential Family) [Alili, 2022]
The exponential family of distributions consists of all distributions with PDF or PMF of the form:

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{30}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are functions that determine the distribution, θ is called the natural parameter and ϕ is called the dispersion parameter.

In addition:

$$\mathbb{E}[Y] = b'(\theta),$$

$$\mathrm{Var}[Y] = a(\phi)b''(\theta),$$

where $b'(\cdot)$ and $b''(\cdot)$ denote the first and second derivates of $b(\cdot)$ respectively.

Of the most common GLMs are logistic linear models, Poisson linear models and of course normal linear models.

Likelihood 2. (Logistic Linear Regression) The logistic linear regression model for M_j is defined as:

$$y|\alpha, \beta_j \sim \text{Bern}(s(\alpha \iota_n + Z_j \beta_j)), \tag{31}$$

where "Bern" represents the Bernoulli distribution and s is the logistic sigmoid function defined as:

$$s(x) = \frac{1}{1 + \exp(-x)}.$$

Likelihood 3. (Poisson Linear Regression) The Poisson linear regression model for M_j is defined as:

$$y|\alpha, \beta_j \sim \text{Poi}(y; e^{\alpha \iota_n + Z_j \beta_j}), \tag{32}$$

where "Poi" represents the Poisson distribution and $e^{\alpha \iota_n + Z_j \beta_j}$ is the rate parameter.

Unlike the normal linear model, when using Zellner's g-prior alongside the independent Jeffreys prior to α and σ^2 defined in (7), the logistic and Poisson models do not give us an analytical form for the marginal likelihoods as normal priors no longer are conjugate [Li and Clyde, 2018]. This adds to the computational challenge as MCMC algorithms would need to sample from both the parameter space and the model space. However, Laplace approximations to the likelihood [Tierney et al., 1989] alongside normal priors such as g-priors can be used to achieve computational efficiency [Li and Clyde, 2018].

The priors on the model parameters in [Li and Clyde, 2018] employ a different type of "centering" of the covariates - one that is induced by the observed information matrix at the maximum likelihood estimator (MLE) of the coefficients. This leads to a g-prior that displays local orthogonality properties at the MLE. They also use a wider class of hyperpriors for g, by considering the class of compound confluent hypergeometric distributions, which contains most hyperpriors used in the literature as special cases. Their results rely on approximations, and their prior structures are data-dependent [Steel, 2020].

For the complementary shrinkage factor $u = 1/(g + 1)$, [Li and Clyde, 2018] argue that

$$p(y|M_j, u) \propto u^{k_j/2} \exp\left(-uS_j/2\right),$$

where S_j is the Wald Statistic (under observed information) for model M_j, and thus argue that a conjugate prior for u should contain the kernel of a truncated Gamma density with the support $u \in [0,1]$.

A generalised Beta distribution, introduced by [Gordy, 1998], and called the Compound Confluent Hypergeometric (CCH) distribution, whose density function contains both Gamma and Beta kernels, is considered.

The Compound Hypergeometric Information Criteria (CHIC) g-prior, encompass several existing mixtures of g-priors as follows: [Li and Clyde, 2018]

Hyper-g Prior [Liang et al., 2008, Cui and George, 2008]

$$u \sim \text{Beta}(\frac{a_h}{2} - 1, 1),$$

with default value $a_h = 3$. When $a_h = 4$, this reduces to a uniform prior on u. $a_h = 2$ corresponds to the Jeffreys' prior on g, which is an improper prior and leads to indeterminate Bayes factors whenever the null model is included in the space of models.

Benchmark Prior [Ley and Steel, 2012]

$$u \sim \text{Beta}(c, c \cdot \max\{n, k^2\}),$$

which induces an approximate prior mean $\mathbb{E}(g) \approx \max\{n, k^2\}$. The recommended parameter value is $c = 0.01$.

Beta-Prime Prior [Maruyama and George, 2011]

$$u \sim \text{Beta}\left(\frac{1}{4}, \frac{n - k_j - 1.5}{2}\right),$$

which is equivalent to a Beta-prime prior on g.

Robust Prior [Bayarri et al., 2012]
The robust prior is a mixture of g-priors with the following hyper prior:

$$p_r(u) = a_r[\rho_r(b_r + n)^{a_r}] \frac{u^{a_r - 1}}{[1 + (b_r - 1)u]^{a_r + 1}} \mathbf{1}_{\{0 < u < \frac{1}{\rho_r(b_r+n)+(1-b_r)}\}},$$

where $a_r > 0$, $b_r > 0$ and $\rho_r \geq b_r/(b_r + n)$.

Hyper-g/n Prior [Liang et al., 2008]

$$p(g) = \frac{a_h - 2}{2n} \left(\frac{1}{1 + g/n} \right)^{a_h/2},$$

where $2 < a_h \leq 4$.

Intrinsic Prior [Berger and Pericchi, 1996, Moreno et al., 1998, Womack et al., 2014] Intrinsic prior is a mixture of g-priors that truncates the support of g. It has the hyper prior:

$$g = \frac{n}{k_j + 1} \cdot \frac{1}{w}, \qquad w \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2} \right).$$

Confluent Hypergeometric (CH) Prior

$$u \sim CH(\frac{a}{2}, \frac{b}{2}, \frac{s}{2}),$$

where "CH" is the Confluent Hypergeometric distribution (defined in [Li and Clyde, 2018]).

Truncated Gamma Prior [Wang and George, 2007, Held et al., 2015]

$$u \sim TG_{(0,1)}(a_t, s_t) \iff p(u) = \frac{s_t^{a_t}}{\gamma(a_t, s_t)} u^{a_t - 1} e^{-s_t u} \mathbf{1}_{\{0<u<1\}},$$

with parameters $a_t, s_t > 0$ and support $[0, 1]$. Here $\gamma(a, s) = \int_0^s t^{a-1} e^{-t} dt$.

[Li and Clyde, 2018] recommend the benchmark beta (see section 2.3.1) and hyper-g/n priors due to their balanced performance in selection and prediction - similar to the findings of [Ley and Steel, 2012] in normal linear models.

6.2 Generalised Linear Models Application

We use the same dataset as in section 5 from [Bruns and Ioannidis, 2020], with the outcome variable transformed to a binary variable, by letting $y = 0$ when GDP growth is lower than the mean in the dataset and $y = 1$ when GDP growth is greater than the mean. This is performed to allow for logistic regression. Using the BAS package from [Clyde, 2021], we fit a logistic regression model using likelihood 2 and the prior structure in (7).

In addition to [Li and Clyde, 2018] recommending the benchmark beta and hyper-g/n priors, they consider the uniform and Beta-Binomial$(1, 1)$ priors over the model space.

We thus reproduce the case of the benchmark beta prior and uniform prior on the model space in [Li and Clyde, 2018] for our modified dataset. The benchmark beta prior has an expected value of $k^2 = 37^2$, and the uniform prior on the model space assigns equal probabilities to all model sizes.

Table 9 gives the posterior inclusion probabilities of the top 7 variables by PIP. In addition, the expected parameter value and the standard deviation is given for each variable. Note the absence of Population in the table; this is perhaps due to the outcome variable losing information of magnitude (e.g. consider the case of China and India) - as it has been transformed to binary values.

Table 9: BMA Output for GLMs - Top 7 Variables by PIP

Explanatory Variable	PIP	$E[\beta_i\|y]$	$\sqrt{\text{Var}(\beta_i\|y)}$
Christianity	0.61768	-2.7967	2.7588
Log GDP Per Capita	0.59223	-2.1367	2.3720
Sub-Saharan Africa	0.55307	-2.7824	175.0977
British Colony	0.33772	1.0910	1.8224
Fertility	0.31609	-0.8032	1.4137
Islam	0.25652	1.3164	2.6745
Life Expectancy	0.25027	0.0897	0.1817

We can also obtain marginal posterior distributions as we did in section 5. Figure 12 illustrates the marginal densities of the two variables with highest PIP: Christianity and

Log GDP Per Capita. The black vertical bar represents the probability the variable is equal to 0.

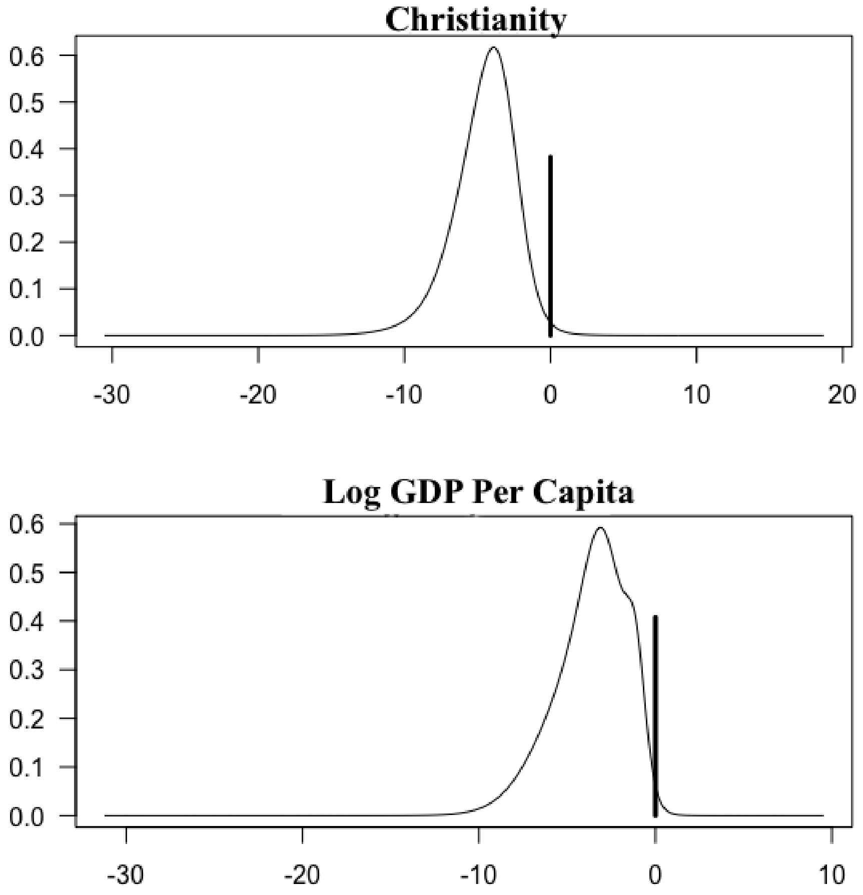


Figure 12: Marginal Posterior Distributions for Christianity and Log GDP Per Capita

Similar to the normal linear setting, our model is fairly confident these variables are to be included with a negative coefficient.

7 Conclusions

In this dissertation, we have explored the theoretical cause and empirical implications of model uncertainty. We have considered methods used in the literature to tackle such uncertainty whilst taking a special focus on Bayesian Model Averaging. We have explained the statistical theory behind BMA, and why this method provides us with a formal framework to implement model averaging. Furthermore, theoretical reasoning has been provided to why some prior choices proposed in the literature can lead to rather different posterior conclusions. Additionally, through MCMC algorithms, we explain why Bayesian methods have become more widely applicable and as flexible as ever, and specifically why they are very important in BMA. Moreover, we have explored various proposals in the literature to implement FMA, and how these methods differ to BMA.

In section 5, we have probed and tested BMA using a real-world recent dataset and measured both its explanatory and predictive capabilities. In particular, we tested BMA against model selection methods such as Bayesian Model Selection, stepwise variable selection and LASSO. Our results indicate a clear advantage of Bayesian methods for model inference through providing marginal distributions for variables, rather than just providing coefficients. Further, in terms of predictive power we have seen that BMA performs as well as LASSO and better than the other considered methods. Lastly in section 5, the importance of prior choices have been highlighted using the same dataset and methodology. There is no one correct prior to use, and the ideal choice is often different based on the goals of the data analysis. However, it is recommended to be aware of certain prior effects and what this can mean for the posterior, and predictive, results.

Throughout this project, the normal linear regression model has been considered primarily, as it is the most common. However, in section 6 we briefly examine BMA in the context of GLMs, and apply a proposal in the literature to a modified version of the dataset used in section 5 to allow for logistic regression.

An area of research that can be developed further is the choice of priors in BMA. In particular, choices of robust priors for both the model space and the parameters. In addition, further work on GLMs is needed to develop a wide framework of models for BMA. Ongoing research into FMA can help provide similar advantages of model averaging in the frequentist framework. [Amini and Parmeter, 2012] compare methods of FMA to BMA and conclude that FMA can provide promising results. In this project, we have not applied FMA to a real-world dataset due to computational restrictions, however methods that can make FMA more computationally accessible have been and continue to be researched in the literature.

Finally, as mentioned in the introduction, despite this project taking a focus on the economic growth context, model uncertainty is a phenomenon present in various disciplines. Results from this project can be used and applied in many fields.

8 References

[Alili, 2022] Alili, L. (2022). St402 risk theory [lecture slides]. University of Warwick.

[Amini and Parmeter, 2012] Amini, S. M. and Parmeter, C. F. (2012). Comparison of model averaging techniques: Assessing growth determinants. *Journal of Applied Econometrics*, 27(5):870–876.

[Bayarri et al., 2012] Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for bayesian model choice with application to variable selection. *The Annals of statistics*, 40(3):1550–1577.

[Berger and Pericchi, 1996] Berger, J. O. and Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

[Brown et al., 1998] Brown, P. J., Vannucci, M., and Fearn, T. (1998). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 12(3):173–182.

[Bruns and Ioannidis, 2020] Bruns, S. B. and Ioannidis, J. P. (2020). Determinants of economic growth: Different time different answer? *Journal of Macroeconomics*, 63:103185.

[Buckland et al., 1997] Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, pages 603–618.

[Clyde, 2021] Clyde, M. (2021). *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. R package version 1.6.0.

[Clyde and George, 2004] Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical science*, 19(1):81–94.

[Cui and George, 2008] Cui, W. and George, E. I. (2008). Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900.

[Doppelhofer and Weeks, 2009] Doppelhofer, G. and Weeks, M. (2009). Jointness of growth determinants. *Journal of Applied Econometrics*, 24(2):209–244.

[Everitt, 2022] Everitt, R. (2022). St420 statistical learning and big data [lecture slides]. University of Warwick.

[Fernandez et al., 2001a] Fernandez, C., Ley, E., and Steel, M. F. (2001a). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

[Fernandez et al., 2001b] Fernandez, C., Ley, E., and Steel, M. F. (2001b). Model uncertainty in cross-country growth regressions. *Journal of applied Econometrics*, 16(5):563–576.

[Foster and George, 1994] Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

[George and McCulloch, 1993] George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

[Glen, 2021] Glen, S. (2021). Likelihood function: Overview / simple definition. `https://www.calculushowto.com/likelihood-function-definition/`. (accessed: 06/12/2021).

[Good, 1952] Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society*, 14:107–114.

[Gordy, 1998] Gordy, M. B. (1998). A generalization of generalized beta distributions.

[Hackenberger, 2019] Hackenberger, B. K. (2019). Bayes or not bayes, is this the question? *Croatian medical journal*, 60(1):50.

[Hansen, 2007] Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

[Hansen and Racine, 2012] Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

[Held et al., 2015] Held, L., Bové, D. S., and Gravestock, I. (2015). Approximate bayesian model selection with the deviance statistic. *Statistical Science*, pages 242–257.

[Houssineau, 2021] Houssineau, J. (2021). St337 / st405 bayesian forecasting and intervention [lecture notes]. University of Warwick.

[Johansen, 2020] Johansen, A. M. (2020). St407 monte carlo methods [lecture notes]. University of Warwick.

[Kass and Wasserman, 1995] Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934.

[Lamnisos et al., 2013] Lamnisos, D., Griffin, J. E., and Steel, M. F. (2013). Adaptive mcˆ 3 and gibbs algorithms for bayesian model averaging in linear regression models. *arXiv preprint arXiv:1306.6028*.

[Lee, 1996] Lee, H. (1996). Model selection for consumer loan application data. *Dept. of Statistics Working Paper*, (650).

[Levine and Renelt, 1992] Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American economic review*, pages 942–963.

[Ley and Steel, 2007] Ley, E. and Steel, M. F. (2007). Jointness in bayesian variable selection with applications to growth regression. *Journal of Macroeconomics*, 29(3):476–493.

[Ley and Steel, 2009] Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of applied econometrics*, 24(4):651–674.

[Ley and Steel, 2012] Ley, E. and Steel, M. F. (2012). Mixtures of g-priors for bayesian model averaging with economic applications. *Journal of Econometrics*, 171(2):251–266.

[Li and Clyde, 2018] Li, Y. and Clyde, M. A. (2018). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, 113(524):1828–1845.

[Liang et al., 2008] Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

[Madigan et al., 1995] Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232.

[Magnus and De Luca, 2016] Magnus, J. R. and De Luca, G. (2016). Weighted-average least squares (wals): a survey. *Journal of Economic Surveys*, 30(1):117–148.

[Magnus et al., 2010] Magnus, J. R., Powell, O., and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of econometrics*, 154(2):139–153.

[Maruyama and George, 2011] Maruyama, Y. and George, E. I. (2011). Fully bayes factors with a generalized g-prior. *The Annals of Statistics*, 39(5):2740–2765.

[Masanjala and Papageorgiou, 2008] Masanjala, W. H. and Papageorgiou, C. (2008). Rough and lonely road to prosperity: a reexamination of the sources of growth in africa using bayesian model averaging. *Journal of applied Econometrics*, 23(5):671–682.

[Moreno et al., 1998] Moreno, E., Bertolino, F., and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association*, 93(444):1451–1460.

[Raftery et al., 1997] Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

[Robert et al., 2007] Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.

[Sala-i Martin, 1997] Sala-i Martin, X. (1997). I just ran two million regressions. *American Economic Review*, 87(2):178–83.

[Sala-i Martin et al., 2004] Sala-i Martin, X., Doppelhofer, G., and Miller, R. I. (2004). Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. *American economic review*, pages 813–835.

[Speagle, 2019] Speagle, J. S. (2019). A conceptual introduction to markov chain monte carlo methods. *arXiv preprint arXiv:1909.12313*.

[Steel, 2020] Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, 58(3):644–719.

[Tierney et al., 1989] Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the american statistical association*, 84(407):710–716.

[Ulam, 1991] Ulam, S. M. (1991). *Adventures of a Mathematician.* Univ of California Press.

[Wang and George, 2007] Wang, X. and George, E. I. (2007). Adaptive bayesian criteria in variable selection for generalized linear models. *Statistica Sinica*, pages 667–690.

[Womack et al., 2014] Womack, A. J., León-Novelo, L., and Casella, G. (2014). Inference from intrinsic bayes' procedures under model selection and uncertainty. *Journal of the American Statistical Association*, 109(507):1040–1053.

[Zellner, 1986] Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.

[Zellner and Siow, 1980] Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, 31(1):585–603.

[Zeugner and Feldkircher, 2015] Zeugner, S. and Feldkircher, M. (2015). Bayesian model averaging employing fixed and flexible priors: The BMS package for R. *Journal of Statistical Software*, 68(4):1–37.

[Zhang et al., 2016] Zhang, X., Yu, D., Zou, G., and Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111(516):1775–1790.

Appendix A Additional Information

Below is the list of countries considered in the dataset used in section 5 and section 6.

Argentina	Australia	Austria	Benin
Bolivia	Brazil	Cameroon	Canada
Central African Rep.	Chile	China	Colombia
Congo	Costa Rica	Denmark	Dominican Rep.
Ecuador	Egypt	El Salvador	Finland
France	Gabon	Ghana	Greece
Guatemala	Honduras	India	Iran
Ireland	Italy	Jamaica	Japan
Jordan	Malaysia	Mali	Mauritania
Mexico	Morocco	Nepal	Netherlands
New Zealand	Niger	Norway	Pakistan
Panama	Paraguay	Peru	Philippines
Portugal	Senegal	South Africa	Spain
Sri Lanka	Sweden	Switzerland	Thailand
Togo	Tunisia	Turkey	United Kingdom
United States	Uruguay	Venezuela	

Figure A.1 illustrates the relationship between the posterior model probabilities and the individual variables. Blue means the variable was included in the model with a positive coefficient and red means the variable was included with a negative coefficient.

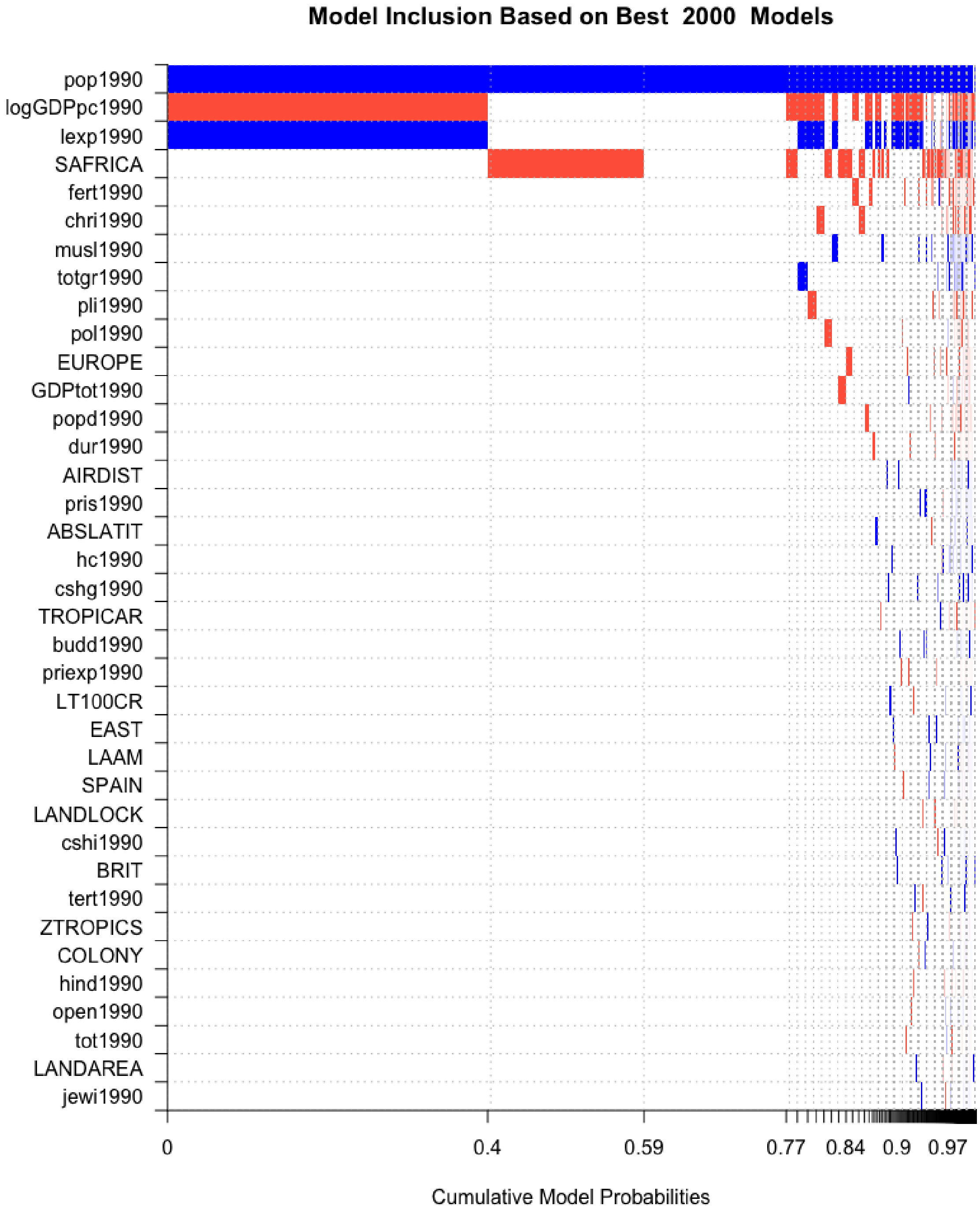


Figure A.1: Best Models and their Corresponding Variables

Similarly, Figure A.2 illustrates the relationship between the log posterior odds and the individual variables in the logistic model.

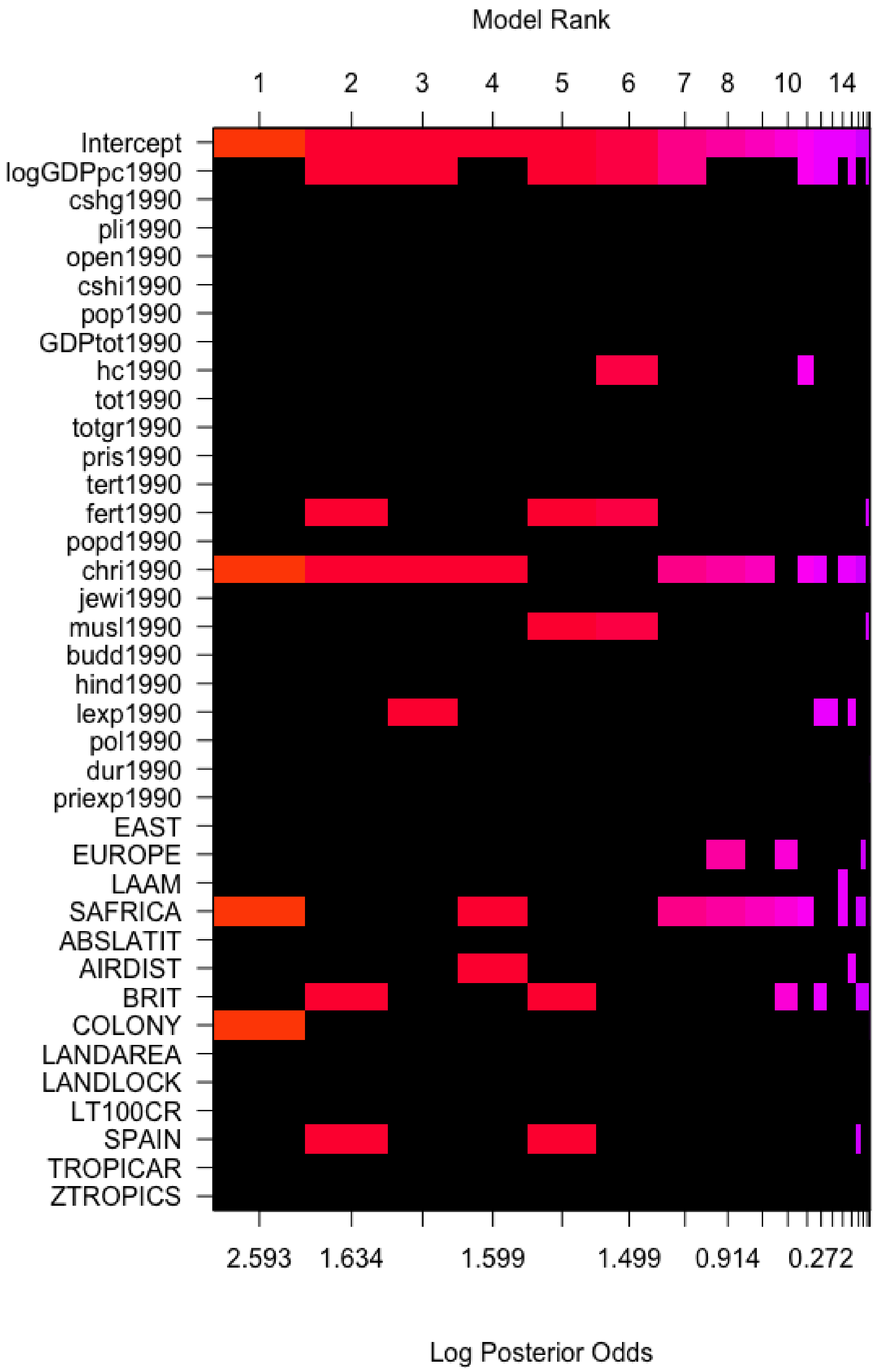


Figure A.2: Variable Inclusion by Model Rank

Appendix B Code Used

Below is the R code used to obtain Figure 2:

```r
#Not an application of Bayes’ rule as such, just an illustration of how 2
   different distributions (prior and likelihood) combine to create another
   distribution (posterior).

N <- seq(-10, 10, 0.01)
likelihood <- dnorm(N, mean = 3, sd = 1)
prior <- dnorm(N, mean = -3, sd = 2)
non.normalised.posterior <- likelihood*prior
require(pracma)
fy <- trapz(N,non.normalised.posterior)
posterior <- non.normalised.posterior/fy

plot(N,posterior,type="l", xlim = c(-10, 10), ylim = c(0, 1), col = "purple",
   lty = 2, xlab = "x", ylab = "Density", lwd = 2, main = "Non-Informative
   Prior")
lines(N, likelihood, type = "l", col = "red", lwd = 2)
lines(N, prior , type = "l", col = "blue", lwd = 2)
legend(x = "topleft", legend = c("Posterior", "Likelihood", "Prior"), col =
   c("purple", "red", "blue"), lty = c(2, 1, 1), lwd = c(2, 2, 2))

likelihood2 <- dnorm(N, mean = 3, sd = 1)
prior2 <- dnorm(N, mean = -3, sd = 0.5)
non.normalised.posterior2 <- likelihood2*prior2
fy2 <- trapz(N,non.normalised.posterior2)
posterior2 <- non.normalised.posterior2/fy2

par(mfrow = c(1, 2))
plot(N,posterior2,type="l", xlim = c(-10, 10), ylim = c(0, 1), col =
   "purple", lty = 2, xlab = "x", ylab = "Density", lwd = 2, main =
   "Informative Prior")
lines(N, likelihood2, type = "l", col = "red", lwd = 2)
lines(N, prior2 , type = "l", col = "blue", lwd = 2)
legend(x = "topright", legend = c("Posterior", "Likelihood", "Prior"), col =
   c("purple", "red", "blue"), lty = c(2, 1, 1), lwd = c(2,2,2))
```

Below is the R code used to obtain Figure 3:

```
m1 <- 7
m2 <- 25
m3 <- 40
k <- 50
ki <- 10
intervals <- seq(0,50, by = 0.01)
pf.odds1 <- function(x){log((m1/(k-m1))^(ki-x))}
pf.odds2 <- function(x){log((m2/(k-m2))^(ki-x))}
pf.odds3 <- function(x){log((m3/(k-m3))^(ki-x))}
curve(pf.odds1, from = 0, to = 50, col = "blue", ylim = c(-70,80), ylab =
    "Log Prior Odds", xlab = "Number of Regressors", main =
    expression(paste("Fixed ", gamma, sep = "")))
lines(intervals, pf.odds2(intervals), col = "purple")
lines(intervals, pf.odds3(intervals), col = "red")
legend("topleft", legend = c("m = 7", "m = 25", "m = 40"), col =
    c("blue","purple","red"), lty=c(1,1,1), cex=0.53)

pr.odds1 <- function(x){
log((gamma(1+ki)/gamma(1+x))*(gamma(((k-m1)/m1)+k-ki)/gamma(((k-m1)/m1)+k-x)))}

pr.odds2 <- function(x){
log((gamma(1+ki)/gamma(1+x))*(gamma(((k-m2)/m2)+k-ki)/gamma(((k-m2)/m2)+k-x)))}

pr.odds3 <- function(x){
log((gamma(1+ki)/gamma(1+x))*(gamma(((k-m3)/m3)+k-ki)/gamma(((k-m3)/m3)+k-x)))}

curve(pr.odds1, from = 0, to = 50, col = "blue", ylab = "Log Prior Odds",
    ylim = c(-30,15),xlab = "Number of Regressors", main =
    expression(paste("Random ", gamma, sep = "")))
lines(intervals, pr.odds2(intervals), col = "purple")
lines(intervals, pr.odds3(intervals), col = "red")
legend("topleft", legend = c("m = 7", "m = 25", "m = 40"), col =
    c("blue","purple","red"), lty=c(1,1,1), cex=0.53)
```

Below is the R code used in section 5:

```
#Section 5
# Load the data from chosen path: load(".../Data and
   code/Data/data20years.RData")
summary(datfinal5)
library(BMS)

#One BMA output for 20 Year Average
dat <- datfinal5
constants_recent <- dat[,5:18]
recent <- dat[,-c(1:18)]
recent <- recent[,seq(7, 7+7*23, 7)]
recent <- data.frame(recent,constants_recent)
set.seed(415)
bma <- bms(recent, burn = 500000, iter = 2000000, g="BRIC", mprior =
   "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")
# Save the object to chosen path: save(bma, file=paste(".../BMA_RUN.RData"))
summary(bma)
topmodels.bma(bma)[, 1:10]
modelprobs <- topmodels.bma(bma)
sum(modelprobs[38,])
sum(modelprobs[1,])
pmp.bma(bma)[1:10, ]
image(bma)
plotConv(bma[1:100], ylab = "Posterior Model Probability")
colSums(pmp.bma(bma))

#Create Table
ores <- matrix()
cbma <- coef(bma, order.by.pip = FALSE)
ores <- data.frame(ores,cbma[,1])
ores <- ores[,-1]
titlenames <- c("GDP per capita", "Government Consumption Share", "Investment
   Price", "Openess",
               "Investment Share", "Population", "Total GDP", "Human
                  Capital","Terms of Trade", "d Terms of Trade",
               "Primary Schooling", "Tertiary Education", "Fertility",
                  "Population Density", "Christianity", "Judaism",
               "Islam", "Buddhism", "Hindu","Life Expectancy", "Polity",
                  "Duration Regime Change", "Primary Exports",
               "East Asia", "Europe", "Latin America", "Sub-Saharan Africa",
                  "Latitude", "Distance to City",
               "British Colony", "Colony", "Land Area", "Landlock", "Land Near
                  Navigable Water", "Spanish Colony",
               "Fraction Tropics", "Tropical Climate Zone")
ores <- data.frame(titlenames, ores)

coef(bma, order.by.pip = TRUE)[,c(2,3)]
round(coef(bma, order.by.pip = TRUE)[,c(2,3)], 7)
coef(bma,exact=TRUE)
```

```
par(mfrow=c(4,2))
d1 <- density(bma, "pop1990", main = "Population (PIP 99.96%)", addons = "me")
abline(v=c(quantile(d1, c(0.025,0.975))[1], quantile(d1, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d2 <- density(bma, "logGDPpc1990", main = "Log GDP Per Capita (PIP 55.52%)",
    addons = "me")
abline(v=c(quantile(d2, c(0.025,0.975))[1], quantile(d2, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d3 <- density(bma, "lexp1990", main = "Life Expectancy (PIP 52.99%)", addons
    = "me")
abline(v=c(quantile(d3, c(0.025,0.975))[1], quantile(d3, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d4 <- density(bma, "SAFRICA", main = "Sub-Saharan Africa (PIP 27.31%)",
    addons = "me")
abline(v=c(quantile(d4, c(0.025,0.975))[1], quantile(d4, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d5 <- density(bma, "chri1990", main = "Christianity (PIP 2.54%)", addons =
    "me")
abline(v=c(quantile(d5, c(0.025,0.975))[1], quantile(d5, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d6 <- density(bma, "fert1990", main = "Fertility (PIP 2.49%)", addons = "me")
abline(v=c(quantile(d6, c(0.025,0.975))[1], quantile(d6, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d7 <- density(bma, "musl1990", main = "Islam (PIP 2.06%)", addons = "me")
abline(v=c(quantile(d7, c(0.025,0.975))[1], quantile(d7, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d8 <- density(bma, "totgr1990", main = "Average Growth Rate of Terms of Trade
    (PIP 1.86%)", addons = "me")
abline(v=c(quantile(d8, c(0.025,0.975))[1], quantile(d8, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
par(mfrow=c(1,1))

plot(bma)
```

```
#Predictive Results
smp_size <- floor(0.75 * nrow(recent))
for (i in 1:20) { #Create 20 data partitions
  dat1 <- recent
  set.seed(i)
  train_ind <- sample(seq_len(nrow(dat1)), size = smp_size)
  assign(paste("train", i, sep = ""), dat1[train_ind,])
  assign(paste("test", i, sep = ""), dat1[-train_ind,])
}

for (i in 1:20) { #Perform BMA on the 20 data partitions
  set.seed(415)
  if(i==1){bma_train1 <- bms(train1, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==2){bma_train2 <- bms(train2, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==3){bma_train3 <- bms(train3, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==4){bma_train4 <- bms(train4, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==5){bma_train5 <- bms(train5, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==6){bma_train6 <- bms(train6, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==7){bma_train7 <- bms(train7, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==8){bma_train8 <- bms(train8, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==9){bma_train9 <- bms(train9, burn = 50000, iter = 1000000, g="BRIC",
      mprior = "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")}
  if(i==10){bma_train10 <- bms(train10, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==11){bma_train11 <- bms(train11, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==12){bma_train12 <- bms(train12, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==13){bma_train13 <- bms(train13, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==14){bma_train14 <- bms(train14, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==15){bma_train15 <- bms(train15, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==16){bma_train16 <- bms(train16, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==17){bma_train17 <- bms(train17, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==18){bma_train18 <- bms(train18, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==19){bma_train19 <- bms(train19, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
  if(i==20){bma_train20 <- bms(train20, burn = 50000, iter = 1000000,
      g="BRIC", mprior = "random", mprior.size = 7, nmodel = 2000, mcmc =
      "bd")}
}

# Save the objects to chosen path: save(bma_train20,
    file=paste(".../bma_train20.RData"))

#Predictive BMA densities
pdens <- pred.density(bma_train1, newdata = test1)
pdens$fit
quantile(pdens, c(0.025, 0.975))
plot(pdens, 16, main = "Predictive Density for Spain (1474 Models)", addons =
    "e")
abline(v=c(quantile(pdens, c(0.025,0.975))[16,1], quantile(pdens,
    c(0.025,0.975))[16,2]), col = "red", lty = 2)
legend("topleft", c("Expected Value", "95% Credible Int."), col =
    c("red","red"), lty = c(1,2), cex = 0.9, bty = "n")

#Compare BMA to other Bayesian Models
bms_best <- as.zlm(bma, 1)
par(mfrow=c(2,3))
b1 <- density(bms_best, "pop1990", main = "BMS: Population", addons = "me")
abline(v=c(quantile(b1, c(0.025,0.975))[1], quantile(b1, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "95% Credible Int."), col = c("red","red"),
    lty = c(1,2), cex = 0.6, bty = "n")
b2 <- density(bms_best, "logGDPpc1990", main = "BMS: Log GDP Per Capita",
    addons = "me")
abline(v=c(quantile(b2, c(0.025,0.975))[1], quantile(b2, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "95% Credible Int."), col = c("red","red"),
    lty = c(1,2), cex = 0.6, bty = "n")
b3 <- density(bms_best, "lexp1990", main = "BMS: Life Expectancy", addons =
    "me")
abline(v=c(quantile(b3, c(0.025,0.975))[1], quantile(b3, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "95% Credible Int."), col = c("red","red"),
    lty = c(1,2), cex = 0.6, bty = "n")
d1 <- density(bma, "pop1990", main = "BMA: Population (PIP 99.96%)", addons =
    "me")
abline(v=c(quantile(d1, c(0.025,0.975))[1], quantile(d1, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d2 <- density(bma, "logGDPpc1990", main = "BMA: Log GDP Per Capita (PIP
    55.52%)", addons = "me")
abline(v=c(quantile(d2, c(0.025,0.975))[1], quantile(d2, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
d3 <- density(bma, "lexp1990", main = "BMA: Life Expectancy (PIP 52.99%)",
    addons = "me")
abline(v=c(quantile(d3, c(0.025,0.975))[1], quantile(d3, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
par(mfrow=c(1,1))

set.seed(415)
for (i in 1:20) {
  assign(paste("pdens_", i, sep = ""), pred.density(get(paste("bma_train", i,
      sep = "")), newdata = get(paste("test", i, sep = ""))))
}
bma_lps <- c()
for (i in 1:20) {
   bma_lps[i] <- lps.bma(get(paste("bma_train", i, sep = "")), realized.y =
       get(paste("test", i, sep = ""))[,1], newdata = get(paste("test", i, sep
       = ""))[,-1])
}

for (i in c(1:5,7,9,11:20)) { #BMA Model on the 6th, 8th and 10th Data
    Partition is just the null model - which is causing issues with the as.zlm
    and pred.density function
  assign(paste("bms_train", i, sep = ""), as.zlm(get(paste("bma_train", i,
      sep = "")), 1))
}
bms_train6 <- zlm(GR1990 ~ 1, data = train6, g = "BRIC") #Best model found
    manually for three null models due to error with as.zlm function
bms_train8 <- zlm(GR1990 ~ 1, data = train8, g = "BRIC")
bms_train10 <- zlm(GR1990 ~ 1, data = train10, g = "BRIC")

for (i in c(1:5,7,9,11:20)) {
  assign(paste("bms_pdens_", i, sep = ""),
      pred.density(get(paste("bms_train", i, sep = "")), newdata =
      data.frame(get(paste("test", i, sep =
      ""))[,c(variable.names(get(paste("bms_train", i, sep = ""))))[-1]])))
}
pred.density(bms_train1, newdata = test1[,c(variable.names(bms_train1))[-1]])

bms_lps <- c() #Assigning LPS values to the null models in the next three
    lines of code
bms_lps[6] <- -mean(dnorm(x = test6$GR1990, mean = bms_train6$coefficients,
    sd =
    sqrt(t(bms_train6$residuals)%*%(bms_train6$residuals)/(length(test6$GR1990)-1)),
    log = TRUE))
bms_lps[8] <- -mean(dnorm(x = test8$GR1990, mean = bms_train8$coefficients,
    sd =
    sqrt(t(bms_train8$residuals)%*%(bms_train8$residuals)/(length(test8$GR1990)-1)),
    log = TRUE))
bms_lps[10] <- -mean(dnorm(x = test10$GR1990, mean =
    bms_train10$coefficients, sd =
    sqrt(t(bms_train10$residuals)%*%(bms_train10$residuals)/(length(test10$GR1990)-1)),
    log = TRUE))
for (i in c(1:5,7,9,11:20)) {
  bms_lps[i] <- lps.bma(get(paste("bms_pdens_", i, sep = "")), realized.y =
      get(paste("test", i, sep = ""))[,1])
}

length((bms_lps - bma_lps)[bms_lps > bma_lps])
c(min(bms_lps), min(bma_lps))
c(mean(bms_lps), mean(bma_lps))
c(max(bms_lps), max(bma_lps))

#Compare BMA to Stepwise
full <- lm(GR1990 ~ ., data = recent)
null <- lm(GR1990 ~ 1, data = recent)
auto.forward <- step(null, scope = list("upper" = full), direction =
    "forward", trace = 0, k = log(63))
auto.backward <- step(full, scope = list("lower" = null), direction =
    "backward", trace = 0, k = log(63))

#Create 20 backward and forward stepwise functions
for (i in 1:20) {
  full_train <- lm(GR1990 ~ ., data = get(paste("train", i, sep = "")))
  null_train <- lm(GR1990 ~ 1, data = get(paste("train", i, sep = "")))
  assign(paste("auto.forward_train", i, sep = ""), step(null_train, scope =
      list("upper" = full_train), direction = "forward", trace = 0, k =
      log(47)))
  assign(paste("auto.backward_train", i, sep = ""), step(full_train, scope =
      list("lower" = null_train), direction = "backward", trace = 0, k =
      log(47)))
}

#LASSO
library(glmnet)
x_vars <- model.matrix(GR1990 ~ ., recent)[,-1]
y_var <- recent$GR1990
lambda_seq <- 10^seq(2, -2, by = -.1)
for (i in 1:20) {
  set.seed(i)
  train_ind <- sample(seq_len(nrow(recent)), size = smp_size)
  cv_output <- cv.glmnet(x_vars[train_ind, ], y_var[train_ind], alpha = 1,
      lambda = lambda_seq, nfolds = 5)
  best_lam <- cv_output$lambda.min
  assign(paste("lasso_best_", i, sep = ""), glmnet(x_vars[train_ind,],
     y_var[train_ind], alpha = 1, lambda = best_lam))
  assign(paste("lasso_pred_", i, sep = ""), predict(get(paste("lasso_best_",
     i, sep = "")), s = best_lam, newx = x_vars[-train_ind,]))
  assign(paste("lasso_train_ind_", i, sep = ""), train_ind)
}

#Squared error loss
loss_bma <- c()
loss_stepwise_forward <- c()
loss_stepwise_backward <- c()
loss_lasso <- c()
for (i in 1:20) {
  loss_bma[i] <- sum((get(paste("test", i, sep = ""))[,1] -
     predict(get(paste("bma_train", i, sep = "")), newdata =
     get(paste("test", i, sep = ""))[,-1]))^2)/length(get(paste("test", i,
     sep = ""))[,1])
  loss_stepwise_forward[i] <- sum((get(paste("test", i, sep = ""))[,1] -
     predict(get(paste("auto.forward_train", i, sep = "")), newdata =
     get(paste("test", i, sep = ""))[,-1]))^2)/length(get(paste("test", i,
     sep = ""))[,1])
  loss_stepwise_backward[i] <- sum((get(paste("test", i, sep = ""))[,1] -
     predict(get(paste("auto.backward_train", i, sep = "")), newdata =
     get(paste("test", i, sep = ""))[,-1]))^2)/length(get(paste("test", i,
     sep = ""))[,1])
  loss_lasso[i] <- sum((get(paste("test", i, sep = ""))[,1] -
     get(paste("lasso_pred_", i, sep = "")))^2)/length(get(paste("test", i,
     sep = ""))[,1])
}

plot(1:20, loss_bma, ylim = c(0,0.002), pch = 16, col = "black", xlab = "Data
    Partition Index", ylab = "Squared Error Loss", main = "BMA vs Stepwise
    Variable Selection vs LASSO", cex = 1.5)
points(1:20, loss_bma, col = "black", type = "l")
points(1:20, loss_stepwise_forward, pch = 17, col = "orange", cex = 1)
points(1:20, loss_stepwise_forward, col = "orange", type = "l")
points(1:20, loss_stepwise_backward, pch = 17, col = "turquoise", cex = 1)
points(1:20, loss_stepwise_backward, col = "turquoise", type = "l")
points(1:20, loss_lasso, pch = 17, col = "pink", cex = 1)
points(1:20, loss_lasso, col = "pink", type = "l")
legend(1, 0.002, legend=c("BMA", "Forward Stepwise", "Backward Stepwise",
    "LASSO"), col=c("black", "orange", "turquoise", "pink"), pch =
    c(16,17,17,17), cex=0.75)

loss_bma - loss_lasso
loss_bma - loss_stepwise_backward
loss_bma - loss_stepwise_forward

#Effect of Prior assumptions
bma_g <- bms(recent, burn = 500000, iter = 2000000, g="UIP", mprior =
    "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")
# Save to chosen path: save(bma_g, file=paste(".../Prior_Assumptions1.RData"))
summary(bma_g) #BMA with g=1/n
plotModelsize(bma_g)
pmp.bma(bma_g)[1:10, ]
colSums(pmp.bma(bma_g))

bma_g2 <- bms(recent, burn = 500000, iter = 2000000, g="hyper=BRIC", mprior =
    "random", mprior.size = 7, nmodel = 2000, mcmc = "bd")
# Save to chosen path: save(bma_g2,
    file=paste(".../Prior_Assumptions2.RData"))
summary(bma_g2) #BMA with g=hyper-g
plotModelsize(bma_g2)
pmp.bma(bma_g2)[1:10, ]
colSums(pmp.bma(bma_g2))

bma_t <- bms(recent, burn = 500000, iter = 2000000, g="BRIC", mprior =
    "fixed", mprior.size = 7, nmodel = 2000, mcmc = "bd")
# Save to chosen path: save(bma_t,
    file=paste("/.../Prior_Assumptions3.RData"))
summary(bma_t) #BMA with Gamma fixed, m = 7
plotModelsize(bma_t)
pmp.bma(bma_t)[1:10, ]
colSums(pmp.bma(bma_t))

bma_t2 <- bms(recent, burn = 500000, iter = 2000000, g="BRIC", mprior =
    "fixed", mprior.size = 18.5, nmodel = 2000, mcmc = "bd")
# Save to chosen path: save(bma_t2,
    file=paste(".../Prior_Assumptions4.RData"))
summary(bma_t2) #BMA with Gamma fixed, m = 18.5
plotModelsize(bma_t2)
pmp.bma(bma_t2)[1:10, ]
colSums(pmp.bma(bma_t2))

bma_tg <- bms(recent, burn = 500000, iter = 2000000, g="UIP", mprior =
    "fixed", mprior.size = 18.5, nmodel = 2000, mcmc = "bd")
# Save to chosen path: save(bma_tg,
    file=paste("/.../Prior_Assumptions5.RData"))
summary(bma_tg) #BMA with Gamma fixed, m = 18.5 and g=1/n
plotModelsize(bma_tg)
pmp.bma(bma_tg)[1:10, ]
colSums(pmp.bma(bma_tg))

par(mfrow = c(3,1))
plotModelsize(bma, sub = "Benchmark Prior")
plotModelsize(bma_g, sub = "Unit Information Prior")
plotModelsize(bma_g2, sub = "Hyper-g Prior")
par(mfrow = c(1,1))

par(mfrow = c(3,1))
plotModelsize(bma, sub = expression(paste("Random ", gamma, sep = "")))
plotModelsize(bma_t, sub = expression(paste("Fixed ", gamma, ", m = 7", sep =
    "")))
plotModelsize(bma_t2, sub = expression(paste("Fixed ", gamma, ", m = 18.5",
    sep = "")))
par(mfrow = c(1,1))

par(mfrow = c(2,2))
g1 <- density(bma_g, "lexp1990", sub = "g = n", addons = "me")
abline(v=c(quantile(g1, c(0.025,0.975))[1], quantile(g1, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
g2 <- density(bma_g2, "lexp1990", sub = "Hyper-g", addons = "me")
abline(v=c(quantile(g2, c(0.025,0.975))[1], quantile(g2, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
t1 <- density(bma_t, "lexp1990", sub = expression(paste("Fixed ", gamma, ", m
    = 7", sep = "")), addons = "me")
abline(v=c(quantile(t1, c(0.025,0.975))[1], quantile(t1, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
t2 <- density(bma_t2, "lexp1990", sub = expression(paste("Fixed ", gamma, ",
    m = 18.5", sep = "")), addons = "me")
abline(v=c(quantile(t2, c(0.025,0.975))[1], quantile(t2, c(0.025,0.975))[2]),
    col = "red", lty = 2)
legend("topleft", c("Cond. EV", "Median", "95% Credible Int."), col =
    c("red","green","red"), lty = c(1,1,2), cex = 0.6, bty = "n")
par(mfrow = c(1,1))
```

Below is the R code used in section 6.2:

```
#GLMs
library(BAS)
recent.glm <- recent #Make outcome variable binary
recent.glm$GR1990 <- ifelse(recent.glm$GR1990 > mean(recent.glm$GR1990),1,0)
bma.glm <- bas.glm(GR1990 ~ ., family = binomial(link = "logit"), data =
   recent.glm, n.models = 20000, betaprior = g.prior(37^2), method = "MCMC",
   MCMC.iterations = 100000000, modelprior=uniform())
# Save to chosen path: save(bma.glm, file=paste(".../GLM.RData"))
summary(bma.glm)
ghat <- bma.glm$shrinkage/(1-bma.glm$shrinkage)
sort(summary(bma.glm)[,1])
glmcoef <- coef(bma.glm)
plot(bma.glm,4)
image(bma.glm, rotate = F)
par(mfrow=c(2,1))
plot(glmcoef, subset = 16, ask = F)
plot(glmcoef, subset = 2, ask = F)
par(mfrow=c(1,1))
```