Bayesian Methods for Fat Data

Gary Koop

January 25, 2016

1 Introduction

Big Data has the potential to revolutionize the way we do econometrics. Big Data comes in two forms: Tall Data where the number of observations is large and Fat Data where the number of variables is large. In this paper, we describe various Bayesian treatments relating to Fat Data. Empirical applications involving Fat Data increasingly arise in many fields in economics, but are particularly common in macroeconomics. In most countries, government statistical agencies collect data on a wide range of macroeconomic variables (e.g. measures of output, capacity, employment and unemployment, prices, wages, housing, inventories and orders, stock prices, interest rates, exchange rates and monetary aggregates).In the US, the Federal Reserve Bank of St. Louis maintains the FRED-MD monthly data base for well over 100 macroeconomic variables from 1960 to the present (see McCracken and Ng, 2015). Many other countries have similar data sets. And, in an increasingly globalized world where economic developments in one country can affect others, the researcher may wish to work with data for several countries. Thus, the researcher may have dozens or hundreds of variables she may wish to include, but only a few hundred observations.

This raises problems for conventional methods of econometric inference. In the presence of Fat Data, simply estimating a model using conventional methods (e.g. non-informative prior Bayesian methods, least squares or maximum likelihood) will typically lead to very imprecise inference (e.g. large posterior variances or wide confidence intervals). In the case where the number of explanatory variables is greater than the number of observations, conventional methods may simply be infeasible. Intuitively, there is simply not enough information in the data to provide precise estimates of the parameters. One solution to this problem might be to select a more parsimonious model using hypothesis testing methods. But, this, too, runs into problems. Such an approach ignores model uncertainty since it assumes the model selected on the basis of hypothesis tests is the one which generated the data. If we have a regression with K potential explanatory variables, then there are 2^K possible restricted models which include some sub-set of the K variables. With Fat Data, K is large, treating one model as if it were "true" and ignoring the huge number of remaining models is problematic. No model selection procedure is perfect, and the researcher is always uncertain about certain about any chosen model. We want a statistical

methodology that reflects this uncertainty. The fact that the selected model has been chosen using hypothesis testing procedures adds weight to the preceding criticism due to the pre-test problem. That is, conventional p-values used for deciding whether to accept or reject a hypothesis are derived assuming a single hypothesis test has been done. If a sequence of hypothesis tests is done (e.g. an initial hypothesis test suggests a variable can be omitted and then additional hypothesis tests are done on a model which omits this variable), then p-values require adjustment. With 2^K potential models and, thus, a huge number of possible tests, the pre-test problem can be serious in Fat Data problems.

A range of methods (not all of them Bayesian), have been developed for working with Fat Data to surmount to preceding problems. In this paper, we discuss some of the major Bayesian approaches. Bayesian approaches have several advantages which make them particularly suitable for Fat Data problems. They allow for the incorporation of prior information which, if available, can help surmount problems caused by an insufficiency of data information. As we shall see, they allow for formal treatment of model uncertainty and, since Bayesian procedures for model selection are quite different from frequentist hypothesis testing procedures, do not suffer from the pre-test problem.

We will discuss a range of Bayesian Fat Data methods in the context of the regression model. But we stress that there are versions of these methods that can also be used for other macroeconomic models such as VARs. We begin by reminding the reader of basic Bayesian results for the Normal linear regression model that the various approaches draw upon. The Normal linear regression model can be written as:

$$y = X\beta + \varepsilon \tag{1}$$

where y is an $N-$vector of dependent variables, X is a $T \times K$ matrix of explanatory variables and ε is an $N-$vector of errors. ε is $N\left(0, h^{-1}I\right)$ where $h = \sigma^{-2}$ is the precision of the error.

The natural conjugate prior is given by:

$$\beta|h \sim N(\underline{\beta}, h^{-1}\underline{V}) \tag{2}$$

and

$$h \sim G(\underline{s}^{-2}, \underline{\nu}), \tag{3}$$

where $N(.,.)$ denotes the Normal distribution and $G(\underline{s}^{-2}, \underline{\nu})$ the Gamma distribution with mean $\underline{s}^{-2}$ and degrees of freedom $\underline{\nu}$. With the natural conjugate prior, the posterior will be:

$$\beta|h \sim N(\overline{\beta}, h^{-1}\overline{V}) \tag{4}$$

and

$$h \sim G(\overline{s}^{-2}, \overline{\nu}), \tag{5}$$

where

$$\overline{V} = \left(\underline{V}^{-1} + X'X\right)^{-1},$$

$$\overline{\beta} = \overline{V}\left(\underline{V}^{-1}\underline{\beta} + X'y\right),$$

$$\overline{\nu} = \underline{\nu} + N$$

and $\overline{s}^{-2}$ is defined implicitly through

$$\overline{\nu}\overline{s}^2 = \underline{\nu}\underline{s}^2 + \left(y - X\overline{\beta}\right)'\left(y - X\overline{\beta}\right) + \left(\overline{\beta} - \underline{\beta}\right)'\overline{V}^{-1}\left(\overline{\beta} - \underline{\beta}\right).$$

An alternative prior is the independent Normal-Gamma prior which is given by

$$\beta \sim N(\underline{\beta}, \underline{V}) \tag{6}$$

and

$$h \sim G(\underline{s}^{-2}, \underline{\nu}). \tag{7}$$

With this prior, the posterior cannot be written in terms of standard densities, but the conditional posteriors (which can be used in an MCMC algorithm) are:

$$\beta|y, h \sim N\left(\overline{\beta}, \overline{V}\right). \tag{8}$$

and

$$h|y, \beta \sim G(\overline{s}^{-2}, \overline{\nu}), \tag{9}$$

where

$$\overline{V} = \left(\underline{V}^{-1} + hX'X\right)^{-1},$$

$$\overline{\beta} = \overline{V}\left(\underline{V}^{-1}\underline{\beta} + hX'y\right),$$

$$\overline{\nu} = N + \underline{\nu}$$

and

$$\overline{s}^2 = \frac{(y - X\beta)'(y - X\beta) + \underline{\nu}\underline{s}^2}{\overline{\nu}}.$$

Throughout this paper, we will use a cross-country growth regression data set to illustrate our methods. This data set is taken from Fernandez, Ley and Steel (2001). Data sets similar to this have been used in numerous papers which investigate the determinants of economic growth. It contains data on GDP growth in $N = 72$ different countries and 41 explanatory variables (and, thus, if we include an intercept $K = 42$). The Fat Data aspect of this application

arises since there are so many potential explanations for growth, but only a fixed number of countries. The dependent variable is average per capita GDP growth for the period 1960-1992. For the sake of brevity, we will not list all of the explanatory variables in this paper (see Fernandez, Ley and Steel, 2001, for a detailed description of all the data). Tables below (e.g. Table 1) provides a list of short form names for all explanatory variables, which should be enough to provide a rough idea of what each explanatory variable is measuring. Each explanatory variable has been standardized by subtracting off its mean and dividing by its standard deviation.

2 Bayesian Model Averaging

2.1 BMA Overview

Bayesian model averaging (BMA) is a general Bayesian concept that is used in a range of empirical contexts, but has proved particularly useful with Fat Data problems. Instead of aiming to select a single model and presenting estimates or forecasts based on it, BMA involves taking a weighted average of estimates or forecasts from all models with weights given by the posterior model probabilities. The theoretical justification for it can be described very simply. Let M_r for $r = 1, .., R$ denotes R different models. The Bayesian treats models as random variables and posterior model probabilities, $p(M_r|y)$, can be defined as being proportion to a prior model probability, $p(M_r)$ times the marginal likelihood, $p(y|M_r)$. If ϕ is a parameter to be estimated (or a function of parameters) or a variable to be forecast, then the rules of probability imply:

$$p(\phi|y) = \sum_{r=1}^{R} p(\phi|y, M_r) p(M_r|y). \tag{10}$$

Thus, the posterior for ϕ is the average of its posterior in each individual model with weights proportional to $p(M_r|y)$. Note that such a strategy allows for a formal treatment of model uncertainty. That is, unlike model selection procedures which choose a single model and proceed as though it were true, (10) explicitly incorporates the fact that we are only $p(M_r|y)$ sure that M_r generated the data.

How these general ideas are operationalized depends on the particular model set-up. Here we describe a common strategy for how BMA is used in regression models with Fat Data.[1] Given an unrestricted regression such as (10), we can define a set of restricted models of interest (often called the model space) as:

$$y = \alpha\iota_N + X_r\beta_r + \varepsilon \tag{11}$$

where ι_N is a $N \times 1$ vector of ones, X_r is a $N \times k_r$ matrix containing some (or

[1] Moral-Benito (2015) surveys these methods including extensions to deal with panel data and endogenous regressors.

all) columns of X. The $N-$vector of errors, ε, is assumed to be $N\left(0_N, h^{-1}I_T\right)$.[2] Note that we are making the common assumption that every model contains and intercept. This is a standard assumption, but can easily be relaxed with minor modifications to the formulae below. Under this assumption, with 2^{K-1} possible subsets of X, there are 2^{K-1} possible choices for X_r and, thus, the number of models is $R = 2^{K-1}$. When working with Fat Data R can be enormous. In our empirical example, $R = 2^{41}$ which raises serious a serious computational hurdle since estimating every model will be impossible (e.g. if each model could be estimated in 0.001 seconds, it would take hundreds of years to estimate all of our models).

We will return to computational issues shortly. But to see how use of BMA surmounts some of the problems of Fat Data regression, note that many of the models in our model space will be parsimonious. In practice, one often finds that BMA attaches most of the weight to these parsimonious models. This is because marginal likelihoods have a strong reward for parsimony. Estimates or forecasts obtained by averaging across many parsimonious models can often be very flexible, able to explain a wide range of behaviours in a way that a single parsimonious model could not. BMA can be an effective way of achieving this goal of combining parsimony with flexibility . Some researchers increase the chances that BMA leads to parsimonious modelling strategies by choosing the prior model probabilities, $p(M_r)$, so that more prior weight is attached to models with fewer explanatory variables. In this section, we will not adopt such a strategy, but it is a simple extension of what we do (see Ley and Steel, 2009).

2.2 BMA Priors

In theory, any prior can be used for each parameter in each model in the model space. In practice, the size of the model space and associated computational concerns suggest that we only consider priors which lead to analytical posterior and predictive results and can be automatically applied to all models. By the latter statement, we mean priors that do not involve hyperparameters that the researcher has to select individually for each model. Since non-informative priors will not lead to valid marginal likelihoods,[3] proper priors are required.[4]

These considerations have led to researchers in this field to use the g-prior. This is a natural conjugate prior as defined in (2) and (3), and, thus, analytical posterior and predictive results exist. The g-prior involves particular, automatic choices for the prior hyperparameters for β_r. In particular, the prior mean is

[2] Formally, we should put r subscripts on each intercept and error precision. However, since these parameters are common to all models and have the same interpretation in all models we simply write them as α and h.

[3] Textbook Bayesian results show that, when comparing models using posterior model probabilities, it is acceptable to use noninformative priors over parameters which are common to all models (e.g. h). However, informative, proper priors have to be used over all other parameters.

[4] Alternatively, non-informaive priors plus approximate posterior model probabilities can be used. For instance, the Bayesian Information Criterion (BIC) can be shown to be asympotically equivalent to the log of the marginal likelihood and, thus, can be used when doing BMA. Such an approach is sometime called Bayesian Averaging of Classical Estimates (BACE).

$$\underline{\beta}_r = 0$$

and prior covariance matrix is $h^{-1}\underline{V}_r$ where

$$\underline{V}_r = \left(gX_r'X_r\right)^{-1},$$

and g is a scalar. The prior mean is chosen to shrink the coefficient towards zero. After all, with Fat Data most of the variables are likely irrelevant so shrinking towards zero is the sensible and parsimonious choice. The form for the prior covariance matrix was suggested in Zellner (1986) where a detailed justification for using this as a benchmark choice for prior is given. The basic idea is that $h^{-1}\left(X_r'X_r\right)^{-1}$ reflects the amount of information in the data for estimating β_r (i.e. under a non-informative prior, it is the posterior covariance matrix of β_r). By having a prior covariance matrix of $h^{-1}\left(gX_r'X_r\right)^{-1}$ we are saying that the information in the prior takes the same form as the data information and g controls the relative strengths of the prior and data information. If $g = 1$, the prior information and data are given equal weight. If $g = 0.01$ then the prior information only receives one per cent of the weight as the data information. Thus, the complicated problem of selecting priors for many parameters in a huge number of models is reduced to the choice of a single, easy-to-interpret scalar: g. There are a few commonly-used rules of thumb for choosing g (see Fernandez and Steel, 2009) or it can be treated as an unknown parameter with its own prior and estimated from the data.

For the error precision and intercept (which are present in every model), it is common to use the standard noninformative priors:

$$p\left(h\right) \propto \frac{1}{h},$$

and:

$$p\left(\alpha\right) \propto 1.$$

2.3 BMA Posterior

The g-prior is a natural conjugate prior and, thus, the posterior will take the form given in (4) and (5). A textbook result for the natural conjugate prior is that the marginal posterior for regression coefficients have a multivariate-t distribution. Using this and integrating out α, the posterior mean for β_r can be shown to be:

$$E\left(\beta_r|y, M_r\right) \equiv \overline{\beta}_r = \overline{V}_r X_r'y,$$

with posterior covariance matrix:

$$var\left(\beta_r|y, M_r\right) = \frac{\overline{\nu}\overline{s}_r^2}{\overline{\nu}-2}\overline{V}_r$$

and $\overline{\nu} = N - 1$ degrees of freedom where

$$\overline{V}_r = \left[(1+g) X_r' X_r \right]^{-1}$$

and

$$\overline{s}_r^2 = \frac{\frac{1}{g+1} y' P_{X_r} y + \frac{g}{g+1} (y - \overline{y}\iota_N)' (y - \overline{y}\iota_N)}{\overline{\nu}},$$

where:

$$P_{X_r} = M_1 - X_r \left(X_r' X_r \right)^{-1} X_r',$$

where

$$M_1 = I_N - \frac{\iota_N \iota_N'}{N}.$$

Using the g-prior, the marginal likelihood for model r is:

$$p\left(y|M_r\right) \propto \left(\frac{g}{g+1} \right)^{\frac{k_r}{2}} \left[\frac{1}{g+1} y' P_{X_r} y + \frac{g}{g+1} (y - \overline{y}\iota_N)' (y - \overline{y}\iota_N) \right]^{-\frac{N-1}{2}}. \tag{12}$$

2.4 BMA Computation

The preceding sub-section provides the posterior and marginal likelihood for each model in our model space. These are the ingredients necessary to do BMA. If the number of models is small, each model can be estimated and its marginal likelihood calculated using the preceding formulae. However, with Fat Data, this is typically impossible. In such cases simulation methods are used. Just as Bayesians use simulation methods (e.g. Gibbs sampling or Metropolis-Hastings algorithms) to learn about the posterior for the parameters, methods can be designed for simulating from the model space to learn about the posterior model probabilities used by BMA. A popular algorithm for simulating from the model space is called Markov Chain Monte Carlo Model Composition, or MC^3, and was first developed in Madigan and York (1995). It is a Metropolis-Hastings algorithm for the model space.

MC^3 produces a number of draws of models which we denote by $M^{(s)}$ for $s = 1, .., S$. At the s^{th} replication, one of M_r for $r = 1, .., R$ is drawn and we call it $M^{(s)}$. Just as with any MCMC algorithm, we can average results (e.g. parameter estimates or forecasts) over the draws. These averages will converge to the true BMA posterior or predictive estimates as $S \to \infty$. For instance, if ϕ is a parameter of interest, then

$$\widehat{\phi} = \frac{1}{S} \sum_{s=1}^{S} E\left(\phi | y, M^{(s)} \right) \tag{13}$$

will converge to $E(\phi|y)$. Similarly, the frequencies with which models are drawn can be used to calculate Bayes factors. For instance, if the MC^3 algorithm draws the model M_i A times and the model M_j B times, then the ratio $\frac{A}{B}$ will converge to the Bayes factor comparing M_i to M_j. In practice, an initial set of burn-in draws should be discarded and standard MCMC diagnostics can be used to select S to ensure the desired level of accuracy. For BMA, a common practice is to choose a set of models (e.g. the 100 models most often drawn by MC^3) and calculate posterior model probabilities in two ways: analytically (using equation 12) and based on the frequencies they are drawn. If the correlation between these two sets of model probabilities is very high, then enough draws have been taken. If not, more are required.

The MC^3 algorithm developed by Madigan and York (1995) is similar to a Random walk Metropolis-Hastings algorithm in that it generates candidate draws that are a step away from the current draw. That is, a candidate model, M^*, is proposed which is drawn randomly (with equal probability) from the set of models including: i) the current model, $M^{(s-1)}$, ii) all models which delete one explanatory variable from $M^{(s-1)}$, and iii) all models which add one explanatory variable to $M^{(s-1)}$. Candidate models are accepted with probability:

$$\alpha\left(M^{(s-1)}, M^*\right) = \min\left[\frac{p(y|M^*)p(M^*)}{p(y|M^{(s-1)})p(M^{(s-1)})}, 1\right].$$

$p(y|M^{(s-1)})$ and $p(y|M^*)$ can be calculated using (12). If a candidate draw, M^*, is accepted then $M^{(s)} = M^*$, else $M^{(s)} = M^{(s-1)}$.

2.5 BMA: Application

In this sub-section we use the BMA methods just described on the cross-country growth regression data set. Fernandez, Ley and Steel (2001) recommend setting $g = \frac{1}{K^2}$ and we also make this choice. The MC^3 algorithm takes $2,200,000$ draws and discards the first $200,000$ as burn-in replications.

Table 1 presents the main results of our BMA exercise. In addition to presenting posterior means and standard deviations of each regression coefficient, the output of the MC^3 algorithm can be used in various ways to present evidence on which models are supported by the data. For instance, the elements in the column of Table 1 labelled "Prob." can be interpreted as the probability that the corresponding explanatory variable should be included. It is calculated as the proportion of models drawn by the MC^3 algorithm which contain the corresponding explanatory variable. Informally, this is a useful diagnostic for deciding whether an individual explanatory variable does have an important role in explaining economic growth. It can be seen that several variables (i.e. Life expectancy, GDP level in 1960, Equipment investment and Fraction Confucian) do have an important role in explaining economic growth. Regardless of which other explanatory variables are included, these variables almost always exhibit strong explanatory power. However, for the remainder of the explanatory variables there is some uncertainty as to whether they have important roles to play

in explaining economic growth. And, for many of the explanatory variables, there is strong evidence that they should not be included.

The next two columns of Table 1 contain posterior means and standard deviations for each regression coefficient, averaged across models. Remember that models where a particular explanatory variable is excluded are interpreted as implying a zero value for its coefficient. Hence, the average in (13) involves some terms where $E\left[\beta_j|y, M^{(s)}\right]$ is calculated and others where the value of zero is used (i.e. for models which do not include x_j the value $\beta_j = 0$ is included in the average). With the exception of the few variables with high BMA posterior probability, most of the posterior means are small relative to their standard deviations. Thus, BMA is indicating a high degree of uncertainty about which factors explain economic growth and the posterior standard deviations reflect this model uncertainty.

The final two columns in Table 1 present results from the single model with highest marginal likelihood. Such a strategy can be called Bayesian Model Selection (BMS). It can be seen that BMS and BMA results are mostly similar to one another. The model selected by BMS is the one containing explanatory variables which BMA is attaching most weight to. However, the BMS posterior standard deviations are smaller. This is due to the fact that BMS ignores model uncertainty: it selects a single model and then proceeds assuming it is the model which generated the data. Since BMS posterior standard deviations only reflect parameter uncertainty (i.e. the uncertainty associated with estimating a parameter in a given model) whereas BMA posterior standard deviation reflect both parameter and model uncertainty, the latter tend to be larger than the former.

Table 1: Point Estimates and Standard Devs of Regression Coefficients (Mean and standard deviations multiplied by 100)					
	BMA			BMS	
Explanatory Variable	Prob.	Mean.	St. Dev.	Mean	St. Dev.
Primary School Enrolment	0.207	0.104	0.234	0.048	0.018
Life expectancy	0.933	0.961	0.392	0.090	0.020
GDP level in 1960	0.999	−1.425	0.278	−1.463	0.193
Fraction GDP in Mining	0.459	0.147	0.181	0.322	0.108
Degree of Capitalism	0.457	0.151	0.183	0.387	0.094
No. Years Open Economy	0.513	0.260	0.283	0.557	0.138
% Pop. Speaking English	0.069	−0.011	0.047	–	–
% Pop. Speak. For. Lang.	0.068	0.012	0.059	–	–
Exchange Rate Distortions	0.082	−0.017	0.070	–	–
Equipment Investment	0.923	0.552	0.236	0.548	0.128
Non-equipment Investment	0.434	0.136	0.174	0.347	0.099
St. Dev. of Black Mkt. Prem.	0.048	−0.006	0.037	–	–
Outward Orientation	0.037	−0.003	0.029	–	–
Black Market Premium	0.179	−0.040	0.097	–	–
Area	0.030	−0.001	0.021	–	–
Latin America	0.215	−0.082	0.191	–	–
Sub-Saharan Africa	0.738	−0.473	0.347	−0.543	0.124
Higher Education Enrolment	0.046	−0.008	0.056	–	–
Public Education Share	0.032	−0.001	0.024	–	–
Revolutions and Coups	0.031	−0.001	0.023	–	–
War	0.075	−0.014	0.062	–	–

Table 1 (continued): Posteror Estimates and Standard Devs of Regression Coefficients					
	Bayesian Model Averaging			Single Best Model	
Explanatory Variable	Prob.	Mean	St. Dev.	Mean	St. Dev.
Political Rights	0.094	−0.028	0.107	–	–
Civil Liberties	0.131	−0.050	0.015	−0.284	0.176
Latitude	0.041	0.001	0.052	–	–
Age	0.085	−0.015	0.058	–	–
British Colony	0.041	−0.003	0.032	–	–
Fraction Buddhist	0.196	0.047	0.109	–	–
Fraction Catholic	0.128	−0.011	0.121	–	–
Fraction Confucian	0.990	0.493	0.127	0.503	0.090
Ethnolinguistic Fractionalization	0.060	0.010	0.056	–	–
French Colony	0.049	0.007	0.040	–	–
Fraction Hindu	0.126	−0.035	0.120	–	–
Fraction Jewish	0.037	−0.002	0.028	–	–
Fraction Muslim	0.640	0.025	0.023	0.295	0.093
Primary Exports	0.100	−0.029	0.105	−0.352	0.136
Fraction Protestant	0.455	−0.143	0.178	−0.277	0.098
Rule of Law	0.489	0.244	0.279	0.563	0.134
Spanish Colony	0.058	0.010	0.068	–	–
Population Growth	0.037	0.005	0.048	–	–
Ratio Workers to Population	0.045	−0.005	0.043	–	–
Size of Labor Force	0.075	0.018	0.097	–	–

The MC^3 algorithm allows for the calculation of posterior model probabilities by simply counting the proportion of draws taken from each model. For the top ten models, the column of Table 2 labelled "$p(M_r|y)$ MC^3 estimate" contains posterior model probabilities calculated in this way. The column labelled "$p(M_r|y)$ Analytical" contains the exact values for the same ten models calculated using (12). It can be seen that the posterior model probability is widely scattered across models with no single model dominating. In fact, the top ten models account for only a little more than 4% of the total posterior model probability. Table 2 indicates there is a great deal of model uncertainty. In fact, BMS is choosing a model with posterior model probability of less than one percent.

The numbers in tables such as this allow for an assessment of the convergence of the MC^3 algorithm. The analytical and numerical posterior model probabilities are slightly different from one another. These differences are small enough for present purposes and we can be confident that the numbers in Table 1 are approximately correct.

Table 2: Posterior Model Probabilities for Top 10 Models		
	$p(M_r\|y)$ Analytical	$p(M_r\|y)$ MC^3 estimate
1	0.0087	0.0089
2	0.0076	0.0077
3	0.0051	0.0050
4	0.0034	0.0035
5	0.0031	0.0032
6	0.0029	0.0029
7	0.0027	0.0025
8	0.0027	0.0027
9	0.0027	0.0026
10	0.0024	0.0022

In this section, we began with a regression with a large number of explanatory variables (41) relative to the number of observations (72) and shown how BMA or BMS can be used to obtain more parsimonious specifications. BMA does this by attaching weight to a large number of more parsimonious models. In fact, the average number of explanatory variables in a regression drawn by MC^3 is 11.4 which is much less than 41. BMS does this by directly selecting a more parsimonious model with 15 explanatory variables. In these algorithms we used a relatively non-informative prior (i.e. the g-prior contains less information than one data point). However, there are other Bayesian methods for Fat Data which directly use more informative priors or allow for the estimation of the degree of information in the prior (i.e. analogous to estimating g) and it is to these we now turn.

3 Variable Selection and Shrinkage Using Hierarchical Priors

3.1 Overview

In general, any sort prior information can be used with Fat Data to overcome the problems causes by an insufficiency of data information. In the regression model, this usually amounts to making suitable choices for $\underline{\beta}$ and $\underline{V}$ in (2) or (6). If such prior information is available, it is desirable to use it. But, given that $\underline{\beta}$ and $\underline{V}$ contain, in total, $K + K \times (K+1)/2$ free parameters, the prior elicitation task can be daunting if K is large. This has led to several simpler priors being proposed which reduce the prior elicitation problem to a much more simple one. The g-prior of the preceding section was one example of such an approach. The ridge regression prior, which sets $\underline{\beta} = 0$ and $\underline{V} = \tau I$ for a scalar τ is another. Posterior analysis using either of these priors can be done in a simple manner using (4) and (5) or (8) and (9) and prior elicitation involves only the elicitation of the scalars g or τ. And, if the researcher does not

wish to subjectively choose g or τ, they can be estimated from the data. If a natural conjugate prior is used, then the marginal likelihood has the analytical form given in (12). The researcher can specify a grid of values for g or τ (e.g. $g = [0.0001, 0.001, 0.01, 0.1, 1.0]$), evaluate the marginal likelihood for each value in the grid and choose the value for g that yields the highest marginal likelihood (averaged over all MC^3 draws). In this way, we can let the data choose the optimal degree of shrinkage.

Such a strategy may sound like it violates a basic tenet of Bayesian econometrics: that the prior should not depend on the data. However, if we interpret the model in a different fashion, it does not. If g or τ are interpreted as unknown parameters having a Uniform prior over the points in the selected grid, then the strategy outlined in the preceding paragraph can be seen to be equivalent to a valid Bayesian analysis involving the estimation of the parameter g (or τ). This is a simple example of a hierarchical prior: where the prior for a parameter (e.g. β) is written in terms of a hyperparameter (e.g. g) which in turn has its own prior (e.g. the Uniform distribution over the selected grid of values for g). Hierarchical priors are widely used in Bayesian econometrics for a range of purposes. In this section, we will discuss how hierarchical priors can be used to deal with Fat Data regression models for achieving prior shrinkage or doing variable selection or doing BMA. There are a rapidly growing variety of such methods and we cannot hope to cover them all in a short paper. Accordingly, we will cover two of the most popular methods: Stochastic search variable selection (SSVS) and the Least Absolute Shrinkage and Selection Operator (LASSO). The reader learning about more approaches is referred to Korobilis (2013) which discusses a range of hierarchical priors which allow for shrinkage of regression coefficients.

3.2 SSVS

3.2.1 SSVS: Theory

We describe an approach to SSVS given in George and McCulloch (1993).[5] It uses a Normal linear regression model with independent Normal-Gamma prior described in (6) and (7) with one alteration to the Normal prior for the regression coefficients. In (6) this prior is simply Normal with a mean and variance chosen by the researcher. The SSVS prior assumes this to be hierarchical. To explain the basic idea of SSVS, suppose we have a simple regression model where β is a scalar. Its SSVS prior is given by:

$$\beta|\gamma \sim (1-\gamma)\, N\left(0, \tau_0^2\right) + \gamma N\left(0, \tau_1^2\right) \tag{14}$$

where $\gamma = 0$ or 1. Thus, if $\gamma = 0$, the prior for β has variance τ_0^2, while if $\gamma = 1$, the prior for β has variance τ_1^2. Since the prior variance controls the amount of prior shrinkage, if τ_0 is small and τ_1 is large then (14) can either produce an extremely tight prior shrinking β to near zero (if $\gamma = 0$), or a relatively non-informative prior which does little shrinkage (if $\gamma = 1$). γ is treated as

[5] The monograph Chipman, George and McCulloch (2001) provides more detail about the practical implementation of SSVS plus describes some related methods.

an unknown parameter and estimated in a data- based fashion. Thus, the data choose whether to select a variable or omit it (in the sense of shrinking its coefficient to be very near zero).[6] The prior for β is hierarchical since it depends on the parameter γ which has its own prior. A Gibbs sampler can be set up which takes draw of γ and, conditional on these, standard posterior results for the independent Normal-Gamma prior given in (8) and (9) can be used to provide draws of β and h. The remainder of this sub-section provides details for how this is done in the regression model with K explanatory variables where β is no longer a scalar.

We will work with a regression model with independent Normal-Gamma prior given in (6) and (7). The SSVS prior relates to the regression coefficients and not h, so we will not discuss (7) other than to say its hyperparameters can be anything. In practice a non-informative prior is often used for h. The prior mean for the regression coefficients is $\underline{\beta} = 0$ so as to shrink coefficients towards zero although other choices could be made without altering the basic theoretical insights of this section. The key aspect of the SSVS is the prior covariance matrix which is set to be:

$$\underline{V} = DD$$

where D is a diagonal matrix[7] with elements

$$d_i = \begin{cases} \tau_{0i} \text{ if } \gamma_i = 0 \\ \tau_{1i} \text{ if } \gamma_i = 1 \end{cases}$$

for $i = 1, .., K$. Note that we now have $\gamma_i \in \{0, 1\}$ for $i = 1, .., K$ indicating whether each variable is included in the regression or not. We also have K small and large prior variances, τ_{0i}^2 and τ_{1i}^2, respectively. These must be selected by the researcher.

How can we select large and small prior variances? We want τ_{0i}^2 to be small enough so that virtually all of the prior probability is attached to the region where β_i, which is the marginal effect of the i^{th} explanatory variable on the dependent variable, is negligible (i.e. it is to all intents and purposes zero). In some cases, the researcher may have enough knowledge about the application at hand to select τ_{0i}^2 in this way. An approximate rule of thumb is that 95% of the probability of a distribution lies within two standard deviations from its mean. In the present case, this means 95% of the prior probability lies in the interval $-2\tau_{0i} \leq \beta_i \leq 2\tau_{0i}$. This may be enough for the researcher to choose τ_{0i}, but not always. For instance, the value $\tau_{0i} = 0.01$ expresses a prior belief that β_i is less than 0.02 in absolute value. Is $\beta_i = 0.02$ a "small" value or not? The answer to this is data dependent (e.g. depends on empirical application at hand and the units the dependent and explanatory variables are measured in). In light of

[6] Some researchers even use so-called "spike and slab" priors where the first element in the prior is a spike at zero (or $\tau_0^2 = 0$) and the coefficient is shrunk to precisely zero if $\gamma = 0$. See Kuo and Mallick (1997).

[7] Some researchers work with $\underline{V} = DRD$ where R is a prior correlation matrix. This is a straightforward extension of the formulae presented here.

such concerns, a practice recommended by many in the field is to choose τ_{0i}^2 and τ_{1i}^2 in a data-based fashion using an initial estimation procedure. For instance, George, Sun and Ni (2008) recommend what they call a default semi-automatic approach which proceeds as follows: First, use ordinary least squares methods in a regression involving all the explanatory variables to produce $\widehat{\sigma}_i$ which is the standard error of β_i. Second, set $\tau_{0i} = \frac{1}{c} \times \widehat{\sigma}_i$ and $\tau_{1i} = c \times \widehat{\sigma}_i$ for some large value for c (e.g. $c = 10$ or 100). George and McCulloch (1993) provide more motivation and explanation for why this is a sensible thing to do. But the basic idea is that $\widehat{\sigma}_i$ provides a rough estimate of the standard deviation of β_i so that the answer to the question: "how do we choose a small value for the prior variance for β_i?" is "we choose one which is small relative to its standard deviation". Typically, one finds that only a rough estimate of what "small" and "large" prior variances are is enough for SSVS to work well. Thus, the default semi-automatic approach is often used in practice.[8] But more sophisticated approaches are possible. For instance, one could estimate c by choosing the value which maximized the marginal likelihood.

The final aspect of prior choice relates to the vector of variable selection indicator variables, $\gamma = (\gamma_1, .., \gamma_K)'$. A common approach is to assume each element of γ has a prior of the form:

$$\begin{array}{l} \Pr(\gamma_i = 1) = \underline{q}_i \\ \Pr(\gamma_i = 0) = 1 - \underline{q}_i \end{array}. \tag{15}$$

A natural default choice is $\underline{q}_j = 0.5$ for all j, implying each coefficient is *a priori* equally likely to be included as excluded.

Bayesian estimation of the Normal linear regression model using this SSVS prior is done using Gibbs sampling. As noted previously, conditional on γ, we know the form of the prior covariance matrix, $\underline{V} = DD$, and can simply use it plugged into the posterior formulae given in (8) and (9). This provides us with the posterior conditionals, $p(\beta|y, h, \gamma)$ and $p(h|y, \beta, \gamma)$, used in the Gibbs sampler. All that is required to complete the Gibbs sampler is a method for drawing γ. It can be shown that the conditional posterior distribution necessary to do this is:

$$\begin{array}{l} \Pr(\gamma_i = 1|y, \gamma) = \overline{q}_i, \\ \Pr(\gamma_i = 0|y, \gamma) = 1 - \overline{q}_j, \end{array} \tag{16}$$

where

$$\overline{q}_j = \frac{\frac{1}{\tau_{1j}} \exp\left(-\frac{\gamma_j^2}{2\tau_{1j}^2}\right) \underline{q}_j}{\frac{1}{\tau_{1j}} \exp\left(-\frac{\gamma_j^2}{2\tau_{1j}^2}\right) \underline{q}_j + \frac{1}{\tau_{0j}} \exp\left(-\frac{\gamma_j^2}{2\tau_{0j}^2}\right) \left(1 - \underline{q}_j\right)}.$$

[8] Note, however, that it can only be used if $K < N$ since it requires OLS estimation which is not possible if the number of explanatory variables exceeds sample size. If $K \geq N$, the researcher may wish to use an informative prior Bayesian approach (e.g. using the ridge regression prior described previously) to obtain this initial estimate.

Thus, Bayesian inference in the regression model using SSVS proceeds by adding a block for drawing γ to a standard Gibbs sampler for the Normal linear regression model with independent Normal-Gamma prior. The output of this posterior simulator can be used in two ways. The most common way is to simply run the Gibbs sampler and average results in the standard way (e.g. take the average of the draws of β_i as the posterior mean of β_i). Note that, since some draws of γ_i will be zero and others one, some draws of β_i will be shrunk to be virtually zero and others will not be. Since we are averaging over draws with $\beta_i = 0$ and draws with β_i being non-zero, this strategy is similar in spirit to BMA. That is, it is averaging over restricted and unrestricted models in a similar fashion as BMA.

Another strategy is to use SSVS to select explanatory variables. This is similar to BMS. A common strategy is to calculate $\Pr(\gamma_i = 1|y)$ using the Gibbs sampler and then select variables with $\Pr(\gamma_i = 1|y) > d$ where d is some threshold for inclusion (e.g. $d = \frac{1}{2}$). The researcher can then use some standard estimation procedure for the Normal linear regression model using only the selected variables.

3.2.2 SSVS: Application

In this sub-section we illustrate use of SSVS methods in our cross-country growth data set. Table 3 presents posterior results using the default semi-automatic prior elicitation approach with $c = 10$. The regression includes an intercept for which we use a relatively non-informative prior. Results are based on $110,000$ draws of which the first $10,000$ are discarded as the burn-in. The final two columns of Table 3, labelled Single Best Model, adopt a strategy where we set $\gamma_i = 1$ if $\Pr(\gamma_i = 1|y) > \frac{1}{2}$ and set $\gamma_i = 0$ otherwise. We then plug the implied values of τ_{0i}^2 or τ_{1i}^2 into $\underline{V}$ and use the MCMC algorithm for the Normal linear regression model with independent Normal-Gamma prior to estimate the model. This is very similar to the variable selection strategy described at the end of the preceding sub-section, but differs in that it shrinks the coefficients on variables which are not selected to be very close to zero instead of being precisely zero. We do this to illustrate how effective SSVS is at shrinking coefficients.

A comparison of the SSVS results in Table 3 with the BMA results in Table 1 reveals a high degree of similarity. These two very different approaches are yielding very similar estimates and standard deviations for β. An empirical researcher reporting these results would come to virtually the same conclusion regardless of whether she was using BMA or SSVS. If we compare variable selection results (i.e. compare the BMS columns in Table 1 to the Single Best Model columns in Table 3), we also find a high degree of similarity. We obtain the same finding that variable selection, since it ignores model uncertainty, leads to estimates which are usually larger in absolute value and are more precise (i.e. posterior standard deviations are smaller). The variables selected by SSVS are mostly the same as those selected by BMS, although there are a few differences. Our implementation of SSVS is selecting 11 variables which is slightly more parsimonious than the 14 selected by BMS.

Table 3: Point Estimates and Standard Devs of Regression Coefficients (Mean and standard deviations multiplied by 100)					
	SSVS			Single Best Model	
Explanatory Variable	$\Pr(\gamma = 1\|y)$	Mean	St. Dev.	Mean	St. Dev.
Primary School Enrolment	0.256	0.111	0.204	2×10^{-5}	0.002
Life expectancy	0.956	0.991	0.365	1.124	0.236
GDP level in 1960	1.000	−1.410	0.286	−1.299	0.202
Fraction GDP in Mining	0.664	0.204	0.179	0.258	0.107
Degree of Capitalism	0.575	0.170	0.176	0.240	0.108
No. Years Open Economy	0.553	0.248	0.267	0.459	0.141
% Pop. Speaking English	0.171	−0.024	0.071	-2×10^{-5}	0.001
% Pop. Speak. For. Lang.	0.174	0.024	0.086	7×10^{-6}	0.001
Exchange Rate Distortions	0.215	−0.038	0.103	-3×10^{-5}	0.001
Equipment Investment	0.917	0.486	0.230	0.538	0.141
Non-equipment Investment	0.584	0.171	0.175	0.282	0.109
St. Dev. of Black Mkt. Prem.	0.138	−0.012	0.054	-2×10^{-5}	0.001
Outward Orientation	0.129	−0.013	0.055	-7×10^{-6}	0.001
Black Market Premium	0.340	−0.068	0.116	-1×10^{-5}	0.001
Area	0.080	−0.001	0.035	3×10^{-6}	0.001
Latin America	0.285	−0.105	0.205	-6×10^{-5}	0.003
Sub-Saharan Africa	0.699	−0.447	0.362	−0.378	0.135
Higher Education Enrolment	0.120	−0.022	0.100	-9×10^{-6}	0.002
Public Education Share	0.119	0.005	0.047	1×10^{-6}	0.001
Revolutions and Coups	0.110	0.002	0.047	-9×10^{-6}	0.001
War	0.204	−0.034	0.094	-2×10^{-5}	0.001

Table 3 (continued): Posteror Estimates and Standard Devs of Regression Coefficients					
	SSVS			Single Best Model	
Explanatory Variable	$\Pr(\gamma = 1\|y)$	Mean	St. Dev.	Mean	St. Dev.
Political Rights	0.130	-0.033	0.121	-1×10^{-4}	0.004
Civil Liberties	0.187	-0.070	0.181	-2×10^{-4}	0.004
Latitude	0.104	0.006	0.086	3×10^{-5}	0.002
Age	0.237	-0.041	0.093	-2×10^{-5}	0.001
British Colony	0.084	-0.005	0.051	-5×10^{-5}	0.002
Fraction Buddhist	0.324	0.076	0.132	3×10^{-5}	0.001
Fraction Catholic	0.216	-0.023	0.158	-2×10^{-5}	0.002
Fraction Confucian	0.972	0.483	0.154	0.542	0.098
Ethnolinguistic Fractionalization	0.141	0.023	0.085	1×10^{-5}	0.002
French Colony	0.138	0.017	0.067	3×10^{-5}	0.001
Fraction Hindu	0.193	-0.068	0.184	-5×10^{-6}	0.003
Fraction Jewish	0.135	-0.008	0.052	-1×10^{-5}	0.001
Fraction Muslim	0.624	0.255	0.241	0.318	0.101
Primary Exports	0.243	-0.073	0.164	-7×10^{-5}	0.002
Fraction Protestant	0.603	-0.189	0.187	-0.276	0.107
Rule of Law	0.485	0.215	0.264	8×10^{-5}	0.002
Spanish Colony	0.129	0.024	0.109	-2×10^{-5}	0.002
Population Growth	0.116	0.017	0.096	3×10^{-6}	0.002
Ratio Workers to Population	0.132	-0.013	0.071	2×10^{-5}	0.001
Size of Labor Force	0.141	0.046	0.167	9×10^{-5}	0.003

3.3 LASSO

3.3.1 Theory

The LASSO was developed as a frequentist shrinkage and variable selection method for Fat Data regression models in Tibsharani (1996). Whereas OLS estimates minimize the sum of squared residuals, LASSO estimates add a penalty term which depends on the magnitude of the regression coefficients. The LASSO minimizes:

$$(y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{k} |\beta_j|$$

where λ is a shrinkage parameter.

It turns out that LASSO estimates can be given a Bayesian interpretation: they are equivalent to Bayesian posterior modes if independent Laplace priors are placed on the regression coefficients. For our purposes, we will not work with the Laplace distribution directly due to the following useful result: the Laplace distribution can be written as a scale mixture of Normals (i.e. a mixture of Normal distributions with different variances). A Laplace prior for a regression

coefficient can be written as:

$$\begin{aligned} \beta_i &\sim N\left(0, h^{-1}\tau_i^2\right), \\ \tau_i^2 &\sim Exp\left(\frac{\lambda^2}{2}\right) \end{aligned} \tag{17}$$

for $i = 1, .., K$ where $Exp(.)$ denotes the exponential distribution.[9] The fact that, conditional on τ_j^2, we have a Normal prior for the regression coefficients can be exploited in the MCMC algorithm. That is, conditional on $\tau = (\tau_1, .., \tau_K)'$, only minor adaptations to (8) and (9) are required to obtain posterior conditionals for β and h. Adding a new block to the MCMC algorithm for drawing τ is all that is required. We can write the LASSO prior covariance matrix as

$$\underline{V} = h^{-1}DD$$

where D is a diagonal matrix with diagonal elements $d_i = \tau_i$ for $i = 1, .., K$.

In one sense, the LASSO is just a different Normal hierarchical prior for the regression coefficients. That is, the SSVS prior is a mixture of two Normal distributions with different variances, the LASSO prior is a scale mixture of Normal distributions. It can be shown (see Park and Casella, 2008) that the LASSO prior should be better at shrinking coefficients with only weak explanatory power towards zero. Since many of the coefficients in Tables 1 and Tables 3 do appear to have only weak explanatory power, this property of the LASSO is potentially useful. This is a point we will investigate in our empirical work using the LASSO.

To complete the model a prior is required for the shrinkage parameter, λ and the error variance, h. For the former, the following is a convenient choice:

$$\lambda^2 \sim G\left(\underline{\mu}_\lambda, \underline{\nu}_\lambda\right).$$

For the latter, it is common to assume a non-informative prior, $p(h) \propto \frac{1}{h}$, and the formulae below make this choice.

For the regression coefficients and error precision, the MCMC algorithm draws from the posterior conditionals $p(\beta|y, h, \tau)$ and $p(h|y, \beta, \tau)$ using formulae which require only slight alterations to the standard results for the Normal linear regression model given in (8) and (9). In particular, $\beta|y, h, \tau$ is $N\left(\overline{\beta}, \overline{V}\right)$ where

$$\overline{\beta} = \left(X'X + (DD)^{-1}\right)^{-1} X'y$$

and

$$\overline{V} = h^{-1}\left(X'X + (DD)^{-1}\right)^{-1}.$$

[9] The reader may wonder why the prior variance is $\sigma^2\tau_j^2$ as opposed to being simply τ_j^2. It can be shown that using the latter can potentially lead to a posterior with more than one mode.

Next we have $h|y,\beta,\tau$ being $G(\overline{s}^{-2},\overline{\nu})$ where

$$\overline{\nu}=N+K$$

and

$$\overline{s}^2=\frac{(y-X\beta)'(y-X\beta)+\beta'(DD)^{-1}\beta}{\overline{\nu}}.$$

The new blocks in the MCMC algorithm relating to the LASSO are for τ and λ. In practice, it is easier to draw from $\frac{1}{\tau_i^2}$ for $i=1,..,K$ as these posterior conditionals can be shown to be independent of one another and each has an inverse Gaussian distribution. The latter distribution, which we denote by $IG(.,.)$, is rarely used in econometrics. However, standard algorithms exist for taking random draws from the inverse Gaussian and so it is straightforward to include it in an MCMC algorithm. In the present context, $p\left(\frac{1}{\tau_i^2}|y,\beta,h,\lambda\right)$ is $IG(\overline{c}_i,\overline{d}_i)$ with

$$\overline{c}_i=\sqrt{\frac{\lambda^2}{h\beta_i^2}}$$

and

$$\overline{d}=\lambda^2.$$

Finally, it is easy to draw from $p\left(\lambda^2|y,\tau\right)$ since this is a $G(\overline{\mu}_\lambda,\overline{\nu}_\lambda)$ distribution with

$$\overline{\nu}_\lambda=\underline{\nu}_\lambda+2K$$

and

$$\overline{\lambda}=\frac{\underline{\nu}_\lambda+2K}{2\sum_{i=1}^{K}\tau_i^2+\frac{\underline{\nu}_\lambda}{\underline{\mu}_\lambda}}.$$

Thus, Bayesian inference using the LASSO prior can be done using MCMC methods which involve a straightforward extension of textbook results for the Normal linear regression model. What we have described in this section is the basic LASSO. There are several other variants of the LASSO that are popular (e.g. the elastic net LASSO). The interested reader is refer Korobilis (2013) for more examples. But the property they have in common is that they can be written in terms of a hierarchical prior which is a mixture of Normal distributions. Hence, MCMC methods very similar to those for the standard LASSO can be used.

3.3.2 LASSO: Application

We continue our empirical analysis of the cross-country growth data set using LASSO methods. For the prior hyperparameters, we use the relatively noninformative choices of $\underline{\mu}_\lambda=0.05$ and $\underline{\nu}_\lambda=1$. The MCMC algorithm is run for $10,000$ burn in draws followed by $100,000$ included draws. Table 4 contains posterior means and standard deviations of the regression coefficients along with the posterior means of the shrinkage parameters, τ_i for $i=1,..,K$. To

help gauge the estimated degree of shrinkage in the LASSO prior, remember that the prior standard deviation for a regression coefficient is $\sigma\tau_i$ and we find $E(\sigma|y) = 0.0071$. The posterior mean for τ_i is given in the table.

Results using the LASSO are similar to those produced using SSVS or BMA. If we use a rule of thumb where posterior means which are two posterior standard deviations from zero are selected as indicating important explanatory variables, then the LASSO is selecting nine explanatory variables. These variables are also selected by SSVS and BMS. It can also be seen that the LASSO is doing a very good job at shrinking unimportant variables in the sense that their coefficients tend to be very small.

Table 4: Posterior Results for Regression Coefficients with LASSO Prior
(Means and standard deviations of regression coeffs multiplied by 100)

Explanatory Variable	$E(\tau_i\|y)$	Posterior Mean	St. Dev.
Primary School Enrolment	0.293	0.237	0.215
Life expectancy	0.932	1.218	0.182
GDP level in 1960	0.901	−1.144	0.109
Fraction GDP in Mining	0.429	0.303	0.058
Degree of Capitalism	0.158	0.094	0.110
No. Years Open Economy	0.578	0.509	0.084
% Pop. Speaking English	4×10^{-4}	-6×10^{-5}	0.003
% Pop. Speak. For. Lang.	0.122	0.069	0.093
Exchange Rate Distortions	6×10^{-4}	-1×10^{-4}	0.004
Equipment Investment	0.581	0.511	0.081
Non-equipment Investment	0.190	0.118	0.124
St. Dev. of Black Mkt. Prem.	5×10^{-4}	-9×10^{-5}	0.003
Outward Orientation	5×10^{-4}	-9×10^{-4}	0.004
Black Market Premium	6×10^{-4}	-9×10^{-5}	0.004
Area	3×10^{-4}	4×10^{-5}	0.001
Latin America	0.005	0.002	0.017
Sub-Saharan Africa	3×10^{-4}	-1×10^{-5}	0.002
Higher Education Enrolment	6×10^{-4}	-1×104	0.005
Public Education Share	3×10^{-4}	2×10^{-5}	0.001
Revolutions and Coups	0.001	3×10^{-4}	0.047
War	5×10^{-4}	1×10^{-4}	0.002

Table 4 (cont.): Posterior Results for Regression Coefficients with LASSO Prior (Means and standard deviations of regression coeffs multiplied by 100)			
Explanatory Variable	$E(\tau_i \mid y)$	Posterior Mean	St. Dev.
Political Rights	5×10^{-4}	3×10^{-5}	0.002
Civil Liberties	3×10^{-4}	5×10^{-5}	0.002
Latitude	7×10^{-4}	2×10^{-4}	0.003
Age	3×10^{-4}	1×10^{-5}	0.001
British Colony	4×10^{-4}	2×10^{-5}	0.001
Fraction Buddhist	0.436	0.314	0.077
Fraction Catholic	0.373	0.253	0.130
Fraction Confucian	0.645	0.617	0.062
Ethnolinguistic Fractionalization	0.001	4×10^{-4}	0.004
French Colony	0.075	0.039	0.071
Fraction Hindu	8×10^{-4}	2×10^{-4}	0.004
Fraction Jewish	6×10^{-4}	1×10^{-4}	0.002
Fraction Muslim	0.671	0.662	0.087
Primary Exports	6×10^{-4}	-6×10^{-5}	0.004
Fraction Protestant	0.002	-9×10^{-4}	0.013
Rule of Law	0.002	8×10^{-4}	0.009
Spanish Colony	0.007	0.003	0.021
Population Growth	0.002	5×10^{-4}	0.007
Ratio Workers to Population	0.001	1×10^{-4}	0.002
Size of Labor Force	0.349	0.217	0.057

References

Chipman, H., George, E. and McCulloch, R. (2001). The Practical Implementation of Bayesian Model Selection. IMS Lecture Notes - Monograph Series Volume 38.

Fernandez, C., E. Ley and Steel, M. (2001). Model uncertainty in cross-country growth regressions. Journal of Applied Econometrics 16, 53-76.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88, 881-889.

George, E., Sun, D. and Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. Journal of Econometrics, 142, 553-580.

Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic regressions with many predictors. International Journal of Forecasting 29, 43-59.

Kuo, L. and Mallick, B. (1997). Variable selection for regression models. Shankya: The Indian Journal of Statistics (Series B), 60, 65-81.

Ley, E. and Steel, M. (2009). On the effect of prior assumptions in Bayesian Model Averaging with applications to growth regression. Journal of Applied Econometrics 24, 651-674.

McCracken, M. and Ng, S. (2015). FRED-MD: A monthly database for macroeconomic research. Federal Reserve Bank of St. Louis, working paper 2015-012A.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. International Statistical Review 63, 215-232.

Moral-Benito, E. (2015) Model averaging in economics: An overview. Journal of Economic Surveys 29, 46-75.

Park, T. and Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association, 103, 681-686.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B, 58, 267-288.

Varian, H. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives 28, 3-28.

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions". In Goel, P.; Zellner, A. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics 6. New York: Elsevier. pp. 233–243