Mixtures of g-priors for Bayesian model averaging with economic applications

Ley, Eduardo and Steel, Mark F. J.

The World Bank, University of Warwick, Statistics Department

22 November 2010

Mixtures of g-priors for Bayesian Model Averaging with Economic Applications

Eduardo Ley
The World Bank, Washington DC, U.S.A.

Mark F.J. Steel
Department of Statistics, University of Warwick, U.K.

Abstract: We examine the issue of variable selection in linear regression modeling, where we have a potentially large amount of possible covariates and economic theory offers insufficient guidance on how to select the appropriate subset. Bayesian Model Averaging presents a formal Bayesian solution to dealing with model uncertainty. Our main interest here is the effect of the prior on the results, such as posterior inclusion probabilities of regressors and predictive performance. We combine a Binomial-Beta prior on model size with a g-prior on the coefficients of each model. In addition, we assign a hyperprior to g, as the choice of g has been found to have a large impact on the results. For the prior on g, we examine the Zellner-Siow prior and a class of Beta shrinkage priors, which covers most choices in the recent literature. We propose a benchmark Beta prior, inspired by earlier findings with fixed g, and show it leads to consistent model selection. Inference is conducted through a Markov chain Monte Carlo sampler over model space and g. We examine the performance of the various priors in the context of simulated and real data. For the latter, we consider two important applications in economics, namely cross-country growth regression and returns to schooling. Recommendations to applied users are provided.

Keywords: Consistency; Model uncertainty; Posterior odds; Prediction; Robustness

JEL Classification System: C11, O47

Correspondence: Mark F.J. Steel, Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom.
Vox: +44 (0) 24 7652 3369
Email: M.F.Steel@stats.warwick.ac.uk

1. Introduction

We focus on problems of variable selection where we have a potentially large amount of covariates in a linear regression context, and the relevant theory does not offer enough guidance on how to select the "appropriate" subset, based on a sample of small or moderate size n. This problem naturally occurs in various applications in economics, such as cross-country growth regression (Brock *et al.*, 2003) or estimating the returns to schooling (Tobias and Li, 2004). Examples of both of these areas of application will be given. As they are quite different in terms of the ratio of observations to potential regressors, n/k, many other econometric scenarios will be covered as intermediate cases.

The use of Bayesian Model Averaging (BMA) provides a natural solution to model uncertainty, based on formal probabilistic reasoning, and it has been shown to lead to better predictions than simply selecting and using one model. Discussions of BMA in linear regression can be found in Raftery *et al.* (1997), Hoeting *et al.* (1999) and Fernández *et al.* (2001a).

We can view the problem of variable selection in regression models as one of inducing sparsity or parsimony and there are two main approaches in the Bayesian literature. One is through the use of shrinkage priors, which goes back to Strawderman (1971) in the context of estimating a multivariate normal mean. These are absolutely continuous priors on all regression coefficients, but are such that some of the regression coefficients will be "close" to zero in the posterior, inducing sparsity (although strictly speaking zero has prior Lebesgue measure zero). The second approach is the one adopted here, where we assign prior point mass at zero for each of the regression coefficients, thus allowing for formal exclusion of covariates and we have to deal with many models that need comparing, which we will typically average over. This was coined the "spike-and-slab" approach in Mitchell and Beauchamp (1988). Even though we use the second approach, some ideas of "shrinkage" will be borrowed from the first literature, appropriately adapted to suit our needs.

In previous studies of variable selection in linear regression using g-priors, it has been noted that the choice of g is crucial for the behaviour of BMA procedures. In addition, the prior on the model space is an important element of the model, particularly in the way it penalizes larger models. If a priori each covariate is included independently with probability θ in the model, the interaction between θ and g was explored in some detail in Ley and Steel (2009). The use of a hierarchical prior on θ was then recommended as a way to make the analysis more robust with respect to prior assumptions on the model space.

In this paper we go one step further and the hierarchical Bayesian model explored here has a hyperprior on θ (which leads to an integral to compute prior model probabilities, which can fortunately be solved analytically) and a hyperprior on g, which leads to an integral for the marginal likelihood that is solved by adding g into the MCMC procedure by an extra Metropolis-within-Gibbs step.

There have been a number of recent proposals for prior distributions to use on g; the paper reviews these approaches and compares them in a common framework of priors that induce a Beta distribution on the shrinkage factor (corresponding to g or to g/n). The one prior that does not fit in this setting is the prior proposed by Zellner and Siow (1980). Based on earlier recommendations for fixed values of g, we propose a benchmark Beta class of priors, and investigate its properties. An added advantage of using random g is that

We are grateful for comments from participants of the first European Seminar on Bayesian Econometrics held at the Erasmus University, Rotterdam (November 4–5, 2010).

the information paradox of Liang *et al.* (2008) can be avoided. In addition, we want the priors to lead to consistency in the sense of Fernández *et al.* (2001a). We also discuss and propose an estimation method for the marginal information that the sample provides on g. This leads quite naturally to estimating the Bayes factors between models with different priors on g or different fixed values of g. We investigate the behaviour of the various priors in BMA with simulated data and various different sets of real data relating to economic applications; two sets of macroeconomic growth data and one data set regarding returns to eduction. We focus mostly on posterior probability on the model that generated the data and the rate of consistency for the simulated data, while we assess prediction performance and compute Bayes factors between priors for the real data. On the basis of both theoretical properties and empirical performance we finally provide recommendations for the applied user.

Section 2 introduces the Bayesian model, whereas Section 3 discusses the hyperpriors on g. The information in the sample regarding g and Bayes factors between model with different priors on g are examined in Section 4. Section 5 briefly mentions computational issues. Applications to simulated and real data follow in Sections 6 and 7, respectively. Finally, conclusions and recommendations are given in Section 8.

2. The Bayesian Model

We adopt a Normal linear regression model for n observations, grouped in a vector y, using an intercept, α, and explanatory variables from a set of k possible regressors in Z. We allow for any subset of the (standardized) variables in Z to appear in the model. This results in a model space of 2^k possible models, which will thus be characterized by the selection of regressors. We call model M_j the model with the $0 \leq k_j \leq k$ regressors grouped in Z_j, leading to

$$y \,|\, \alpha, \beta_j, \sigma, M_j \sim \mathrm{N}(\alpha \iota_n + Z_j \beta_j, \sigma^2 I), \tag{1}$$

where ι_n is a vector of n ones, $\beta_j \in \Re^{k_j}$ groups the relevant regression coefficients and $\sigma \in \Re_+$ is a scale parameter.

For the parameters in a given model M_j, we follow Fernández *et al.* (2001a) and adopt a combination of a "non-informative" improper prior on the common intercept and scale and a so-called g-prior (Zellner, 1986) on the regression coefficients, leading to the prior density

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

where $f_N^q(w \,|\, m, V)$ denotes the density function of a q-dimensional Normal distribution on w with mean m and covariance matrix V. The regression coefficients not appearing in M_j are exactly zero, represented by a prior point mass at zero (this is the "spike-and-slab" idea). Of course, we need a proper prior on β_j in (2), as an improper prior would not lead to meaningful Bayes factors. The so-called "benchmark" prior structure in (2), sometimes with small variations, is shared by most of the recent literature on covariate selection in linear models—see, *e.g.*, Clyde and George (2004) for a survey.

One advantage of this prior is that we only need to choose a single scalar parameter g.[1] A large fraction of the literature in this area has dealt with this choice of g, and it is clear that posterior and predictive inference critically depends on g; see Ley and Steel (2009), and Eicher *et al.* (2010) for examples of this in the context of growth regressions. Popular

[1] There is one difference with respect to the notation in Fernández *et al.* (2001a); in line with most of the literature, in this paper g denotes a variance factor rather than a precision factor.

values for g in the literature are to take $g = n$, which corresponds to assigning the same amount of information to the conditional prior of β as is contained in one observation—the so-called "unit information prior" of Kass and Wasserman (1995), also favoured in Eicher *et al.* (2010)—or to take $g = k^2$ as suggested by the Risk Inflation Citerion (RIC) of Foster and George (1994). Fernández *et al.* (2001a) recommend the "benchmark" choice of $g = \max\{n, k^2\}$. To avoid having to choose a particular value, we will put a hyperprior on g, allowing for the data to influence the inference on the now random g. This makes the analysis more robust with respect to the assumptions on g and has also been used, among others, in Liang *et al.* (2008), Bottolo and Richardson (2008), Cui and George (2008), and Feldkircher and Zeugner (2009).

Thus, denoting by $p(g \mid M_j)$ the prior for $g > 0$ (which could depend on M_j), we have the following prior for all parameters in M_j

$$p(\alpha, \beta_j, \sigma, g \mid M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j \mid 0, \sigma^2 g (Z_j' Z_j)^{-1}) p(g \mid M_j). \tag{3}$$

In other words, we can interpret the prior on g as a way of extending the normal prior in (2) to a scale mixture of normals, which has substantially more flexible tails—see Andrews and Mallows (1974), and Fernández and Steel (2000).

Now we can write the marginal likelihood as the following (with a proportionality constant that is the same for all models, including the null model):

$$l_y(M_j) \propto \int_0^\infty (1+g)^{\frac{n-1-k_j}{2}} [1 + g(1 - R_j^2)]^{-\frac{n-1}{2}} p(g \mid M_j) \mathrm{d}g, \tag{4}$$

where R_j^2 is the usual coefficient of determination for model M_j; $R_j^2 = 1 - y' Q_{X_j} y / (y - \bar{y}\iota_n)'(y - \bar{y}\iota_n)$, where $Q_A = [I - A(A'A)^{-1}A']$, and $X_j = (\iota_n \; Z_j)$ is the design matrix of model M_j, which we assume to be of full column rank. This marginalised likelihood (4) is simply the sampling density integrated out with the prior, and is of critical importance as the ratio of marginal likelihoods of any two models is the Bayes factor between these models.

Note that the prior on g needs to be proper, as the null model does not involve g and an improper prior on g would thus lead to arbitrary Bayes factors versus the null model. An important quantity in evaluating the properties of priors on g will be the shrinkage factor, which is $\delta = g/(1+g)$. The posterior mean of β_j for each given model will be the OLS solution times this shrinkage factor.

For constructing the prior over model space, we assume that each potential regressor is independently included in the model with probability θ. As in Brown *et al.* (1998), Nott and Kohn (2005) and Ley and Steel (2009), we also consider putting a hyperprior on θ, which is then assigned a Beta distribution. Ley and Steel (2009) illustrate that such a random θ approach renders the analysis much more robust with respect to prior assumptions. They advocate the parameterization

$$\theta \sim \mathrm{Be}(1, (k-m)/m), \tag{5}$$

where m is then the prior mean model size. The resulting hierarchical prior over model space is less restrictive than the one with fixed θ and the choice of m is shown in Ley and Steel (2009) not to matter too much in practice. The prior on model size induced by (5) is a Binomial-Beta distribution—see, *e.g.*, Bernardo and Smith (1994, p. 117).

3. Hyperpriors for g

3.1. Zellner-Siow Prior

Inspired by Jeffreys' (1961) arguments for using Cauchy priors in model comparison problems related to a univariate normal mean, Zellner and Siow (1980) proposed the use of multivariate Cauchy priors in regression problems. As is well known, and mentioned in Liang *et al.* (2008), Student-t form priors (like the Cauchy) can easily be expressed as scale mixtures of normals and thus, the Zellner-Siow prior corresponds to a particular choice of $p(g)$ in (3). In particular, Zellner and Siow implicitly propose an inverted Gamma distribution with parameters $\frac{1}{2}$ and $\frac{n}{2}$, leading to

$$p(g) = \frac{\sqrt{\frac{n}{2}}}{\Gamma(\frac{1}{2})}\, g^{-3/2} \exp\left(-\frac{n}{2g}\right).$$

This implies the following distribution for the shrinkage factor $\delta = g/(1+g)$:

$$p(\delta) = \frac{\sqrt{\frac{n}{2}}}{\Gamma(\frac{1}{2})}\, \delta^{-3/2}(1-\delta)^{-1/2} \exp\left(-\frac{n(1-\delta)}{2\delta}\right). \tag{6}$$

3.2. Beta Shrinkage Priors

A relatively large number of priors in the literature, in fact, imply a Beta prior distribution for the shrinkage factor δ. A Beta(b, c) prior on the shrinkage factor induces the following prior on g:

$$p(g) = \frac{\Gamma(b+c)}{\Gamma(b)\Gamma(c)}\, g^{b-1}(1+g)^{-(b+c)}, \tag{7}$$

which is called an inverted Beta distribution in Zellner (1971, p. 375) and is also known as a Gamma-Gamma distribution (Bernardo and Smith, 1994, p. 120) in the statistics literature. This has the following properties:

$$\begin{cases} \mathrm{E}\,[g] = \frac{b}{c-1} & \text{provided } c > 1 \\ \mathrm{Var}\,[g] = \frac{b(b+c-1)}{(c-1)^2(c-2)} & \text{provided } c > 2 \end{cases}$$

and has a mode equal to $(b-1)/(c+1)$ provided $b > 1$. This inverted Beta prior on g leads to the following prior on the regression coefficients, marginalised over g:

$$p(\beta_j \mid M_j, \sigma) = \frac{\Gamma(b+c)\Gamma\left(c+\frac{k_j}{2}\right)|Z_j'Z_j|^{1/2}}{\Gamma(b)\Gamma(c)(2\pi)^{k_j/2}\sigma^{k_j}} \cdot \Psi\left(c+\frac{k_j}{2},\ \frac{k_j}{2}-b+1;\ \frac{\beta_j' Z_j' Z_j \beta_j}{2\sigma^2}\right),$$

where Ψ denotes the confluent hypergeometric function (Gradshteyn and Ryzhik, 1994, p. 1085). Note that these Beta shrinkage priors have a density in the right tail that behaves like $g^{-(1+c)}$, thus leading to very fat tails for small values of c.

The so-called hyper-g prior, proposed by Liang *et al.* (2008), corresponds to $b = 1$ and $c = \frac{a}{2} - 1$; with $a > 2$ to ensure a proper prior. This class includes priors used by Strawderman (1971) for the normal means problem. Cui and George (2008) propose to use $a = 4$ in the context of model selection with known σ. Bottolo and Richardson (2008) adopt a hyper-g prior with $a = 2$, but make it proper by truncating the right tail at

$\max\{n, k^2\}$, which is the benchmark value for g proposed by Fernández *et al.* (2001a). Feldkircher and Zeugner (2009) propose a hyper-g prior with a value of a that leads to the same mean shrinkage factor as the unit information prior or the RIC prior. These hyper-g priors have a finite nonzero limit as $g \to 0$.

Another prior corresponding to a Beta shrinkage prior is the horseshoe prior of Carvalho *et al.* (2010), where both b and c are taken to be $\frac{1}{2}$. The shrinkage factor thus has a U-shaped prior where the spike around zero corresponds to very strong shrinkage and induces zero coefficients in their shrinkage prior framework, and the spike around one describes the signal.

Whereas the horseshoe prior was explicitly developed for a different setting (shrinkage priors rather than our spike-and-slab framework), and the hyper-g has roots in the shrinkage literature, the latter prior has been proposed explicitly for problems where we do have prior point masses to deal with formal exclusion of regressors—in Liang *et al.* (2008), Cui and George (2008), and Bottolo and Richardson (2008). However, in our spike-and-slab case, we do not need to rely on shrinkage to exclude regressors: this is formally allowed for by the prior on the model space. Thus, we would expect that we don't really need such a large prior mass around small values of the shrinkage factor δ. Nevertheless, the horseshoe prior has an accumulation of mass towards zero for both δ and g. In terms of g, both the horseshoe prior and the hyper-g prior always decrease in g, and the horseshoe prior even has an asymptote at zero. A graphical illustration of the various priors is provided in Figure 2, discussed later.

One way to adapt these priors (partly) motivated by the shrinkage literature to our current setting, is to realize that typically the shrinkage priors are not used in a g-prior (where the conditional covariance of the regression coefficients is proportional to the inverse information matrix) but in a ridge-type prior structure (where this conditional covariance is proportional to the identity matrix). In the g-prior framework used here, we need to account for the fact that information accrues with sample size and the inverse information matrix is of order $1/n$. Thus, a fair comparison with a shrinkage ridge-type prior setup would perhaps be to apply the shrinkage prior to g/n rather than g. Equivalently, we then use a Beta(b, c) prior on $g/(n+g)$. Starting from the hyper-g prior, this leads directly to the hyper-g/n prior of Liang *et al.* (2008). Similarly, we shall denote the prior thus derived from the horseshoe prior by horseshoe/n prior. Of course, this does not change the fact that these priors are always decreasing in g, but it does move some mass towards larger values of g, and makes the right tail of the horseshoe/n prior on δ much thicker than the left (see Figure 2). In addition, it solves an inconsistency problem. As shown in Liang *et al.* (2008), the Zellner-Siow prior and the hyper-g/n prior are consistent in the sense described in Subsection 3.4. This is also the case for the horseshoe/n prior, but not for any prior on g that does not depend on n. Subsection 3.4 presents more results in this respect.

Maruyama and George (2010) propose to choose $b + c = (n - k_j - 1)/2$ and $c < 1/2$ in (6). This choice is motivated by the fact that the integral in (4) then has a simple analytic solution and thus Bayes factors can be computed as easily as in the case with fixed g. In particular, we obtain (up to a common proportionality constant)

$$l_y(M_j) \propto \Gamma\left(\frac{n-k_j-1}{2} - c\right)\Gamma\left(c + \frac{k_j}{2}\right)(1-R_j^2)^{c-\frac{n-k_j-1}{2}},$$

whenever $c < (n - k_j - 1)/2$. As a default value, Maruyama and George (2010) propose to take $c = 1/4$. However, note that this choice implies that the prior on g depends on the

model we are considering (through the model size). Formally, this is allowed, but it may make it slightly harder to interpret the role of g. Maruyama and George (2010) show that consistency holds with this prior.

Extending the prior for robust estimation of Berger (1985), Forte *et al.* (2010) effectively propose the use of a truncated version of (7) with $b = 1$ and $c = 1/2$, while truncating g to be greater than $(n + 1)/(k_j + 3) - 1$. They prove consistency of the resulting model choice procedure and provide a closed-form expression for the Bayes factor. Through the truncation point, the prior on g, again, becomes model-specific.

From fixed g analyses with growth data (Eicher *et al.*, 2010; Ley and Steel, 2009) it seems that rather large values of g could be preferable. In particular, values like $g = n$ or even $g = k^2$ are the most used in this literature. With the hyper-g we can only assign large prior mass to regions with high g by taking a very close to 2, which gives us a fat tail for g, but does not really change the shape of the prior. In order to propose an alternative class of priors in the next subsection, we initially focus on the shrinkage factor, δ.

3.3. A Benchmark Beta Prior

If we start from a Beta(b, c) prior on the shrinkage factor $\delta = g/(1 + g)$, we can base our proposal on ensuring that prior moments are reasonable (they always exist, as we are dealing with a finite support). In particular, let us set the mean shrinkage factor equal to the one that corresponds to the Fernández *et al.* (2001a) recommendation $g = \max\{n, k^2\}$. This fixes one parameter as a function of the other, as then $b/c = \max\{n, k^2\}$. The second parameter can then be chosen by considering the spread around this mean:

$$\text{Var}\left[\delta\right] = \frac{d(1-d)}{1 + c(1 + \max\{n, k^2\})},$$

where

$$d = \frac{\max\{n, k^2\}}{1 + \max\{n, k^2\}}$$

is the chosen prior mean of the shrinkage factor, and thus larger c corresponds to a tighter prior around this mean.

For example, for the cross-country growth data used in Fernández *et al.* (2001b) with $k = 41$ and $n = 72$, the value $c = 1$ would correspond to a prior standard deviation of the shrinkage factor of 0.0006 (approximately equal to $1 - d$) and the corresponding $b = 1681$. For $c = 1/1000$ we obtain a prior standard deviation of 0.015 (approximately 25 times $1 - d$) and this corresponds to $b = 1.681$. For likely choices of c (say, $c = 1, 0.1, 0.01$) and $\max\{n, k^2\} > 100$ (as for typical datasets in econometrics) we have no prior moments for g, but we do have an interior mode equal to

$$\text{Mode}\left[g\right] = \frac{\max\{n, k^2\} - 1/c}{1 + 1/c},$$

which is approximately $\max\{n, k^2\}/2$ for $c = 1$ and is a lot smaller for the smaller values of c.

Figure 1 shows three prior densities (expressed in logs to make matters easier to visualize) of g and δ from this class of benchmark priors. The figure illustrates the effect of c: For $c = 1$ the density is most concentrated while for the smallest c the prior assigns a lot of the

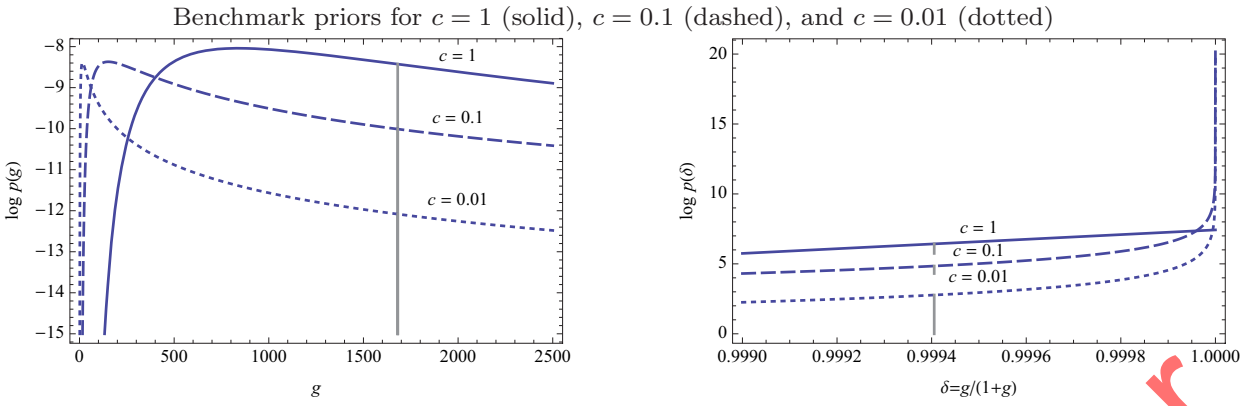


Fig. 1. Beta Benchmark Priors for g and δ ($n = 72$, $k = 41$)
(Vertical lines correspond to $g = \max\{n, k^2\}$)

mass to the far right tail of g (for example, $P(g > 5,000) = 0.95$ for $c = 0.01$). Notably, all priors for g have an interior mode. As $c \to \infty$ the benchmark Beta prior tends in the limit to the case with fixed $g = \max\{n, k^2\}$.

As a special case of the benchmark Beta class, we can get a single hyper-g prior by choosing $c = 1/\max\{n, k^2\}$. This would lead to a hyper-g prior with

$$a = 2 \times \frac{\max\{n, k^2\} + 1}{\max\{n, k^2\}}.$$

In fact, this would effectively correspond to taking the prior setting with the smallest a of the two proposals in Feldkircher and Zeugner (2009). Note that even though the mean shrinkage factor is the same, this is a very different prior, where we achieve the mean shrinkage by a very small c (with fixed $b = 1$) rather than the benchmark choice of a very large b (with fixed c, chosen to give a reasonable prior spread of δ).

Figure 2 helps us understand the differences. It plots the log densities of g and δ for all the priors discussed here (using $n = 72$ and $k = 41$ as in the first growth dataset in Section 7). Interior modes for g occur for the Zellner-Siow, Maruyama-George, Forte *et al.*[2] and benchmark priors. All other priors tend to a positive constant or infinity[3] as $g \to 0$. In order to get the same mean shrinkage factor as the benchmark Beta prior, the Feldkircher-Zeugner priors thus need to decrease very rapidly with g in order to compensate for the mass close to $g = 0$ and then have a very fat far right tail (almost of the order $1/g$). As a consequence, the prior probability assigned by the Feldkircher-Zeugner priors to $\delta > d$ is large (over 0.9 when $a = 2 + 2/n$ and 0.9986 for $a = 2 + 2/k^2$). For the benchmark prior this probability decreases from 0.96 to 0.63 as c ranges from 0.01 to 1. Particularly small prior probabilities for large values of δ are associated with the Bottolo-Richardson prior (where it is 0 due to the truncation) and the hyper-g ($a = 3, 4$) and horseshoe priors, which have relatively thin right tails for g. The Zellner-Siow prior for g has the same right tail behaviour as the hyper-g with $a = 3$ and the horseshoe prior and also leads to relatively small mass on $\delta > d$ (0.16). The Maruyama-George and Forte *et al.* priors are intermediate cases in this respect. Remember also that the latter two priors depend on

[2] Note also that the Forte *et al.* prior for g is truncated away from zero, in that $g > ((n+1)/(k+3)) - 1$. In addition, this prior, computed as in (8), is multimodal, with the truncation leading to a saw-tooth effect for small and moderate values of g (the last discontinuity is at $g = (n-2)/3$).

[3] For the horseshoe and horseshoe/n priors.

the model M_j through its model size, k_j. Thus, denoting model size by W, we compute the marginal prior for g as follows:

$$p(g) = \sum_{w=0}^{k} p(g \mid w) P(W = w), \tag{8}$$

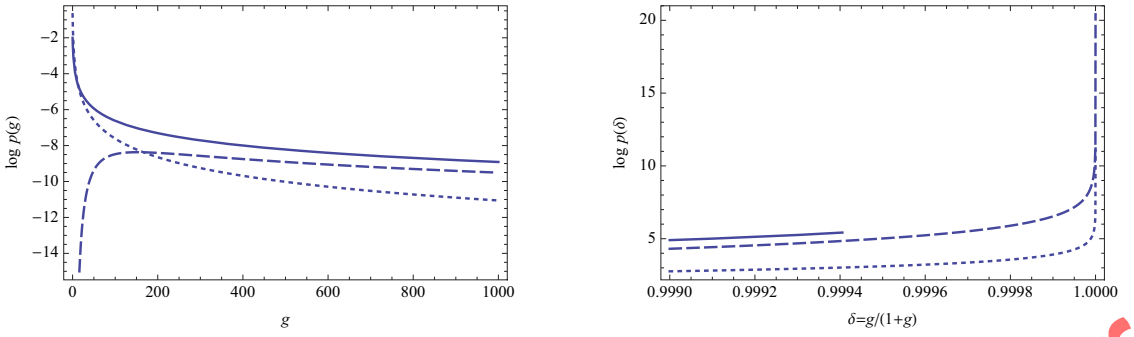
where $P(W = w)$ is the probability mass function of the Binomial(k, θ) (for fixed θ) or Binomial-Beta distribution (for random θ, and as given in Ley and Steel, 2009). Figure 2 presents the marginal Maruyama-George and Forte *et al.* priors in the situation where we use $m = 7$ for the random θ case.

Table 1 summarizes the definition and some of the main properties of the different prior distributions. The names of the priors are as in the text, except for "F-Z," which denotes the Feldkircher-Zeugner prior. The numbers refer to prior numbers as used later in the empirical sections. The column "truncated" indicates any truncation of the support for g and the column "model spec." indicates whether the prior is specific to any given model. In the column "mean" we report whether the prior mean exists (and if so, under which conditions). The availability of analytical expressions for the Bayes factors is indicated in "BF", while the last column "cons." refers to consistency as discussed in the next subsection. Finally, note that if the prior mean of g does not exist, neither will the posterior mean. The reason for that is simply that the posterior equals the prior for the null model (since that model does not involve g). Thus, even if the null model has only a very small posterior probability, the latter is not zero and the overall posterior mean of g will be infinite.

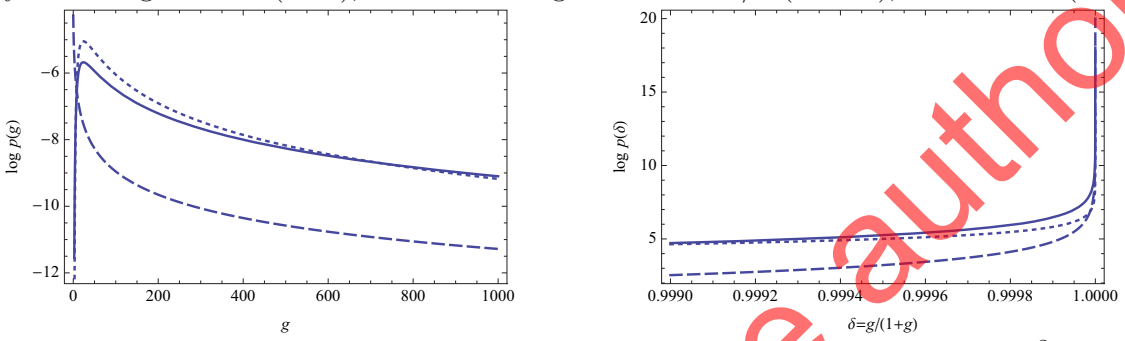
Table 1. Summary of prior distributions on g

		b	c	truncated	model spec.	mean	BF	cons.
		Beta shrinkage priors, $\delta \sim \text{Be}(b, c)$						
1–3	Benchmark	$c \max\{n, k^2\}$	c	no	no	$c > 1$	no	yes
4	Maruyama-George	$\frac{n-k_j-1}{2} - c$	$< 1/2$	no	yes	no	yes	yes
5	Bottolo-Richardson	1	0	$g < \max\{n, k^2\}$	no	yes	no	yes
6–7	Hyper-g	1	$(a/2) - 1$	no	no	$a > 4$	no	no
8	F-Z, $a = 2 + 2/n$	1	$1/n$	no	no	no	no	yes
9	F-Z, $a = 2 + 2/k^2$	1	$1/k^2$	no	no	no	no	no
10	Horseshoe	1/2	1/2	no	no	no	no	no
11	Forte *et al.*	1	1/2	$g > \frac{n+1}{k_j+3} - 1$	yes	no	yes	yes
		Beta shrinkage/n priors, $\frac{g}{n+g} \sim \text{Be}(b, c)$						
		b	c	truncated	model spec.	mean	anal.	cons.
12	Hyper-g/n	1	$(a/2) - 1$	no	no	$a > 4$	no	yes
13	Horseshoe/n	1/2	1/2	no	no	no	no	yes
		Inverted Gamma priors, $g \sim \text{IG}(b, c)$						
		b	c	truncated	model spec.	mean	anal.	cons.
14	Zellner-Siow	1/2	$n/2$	no	no	no	no	yes

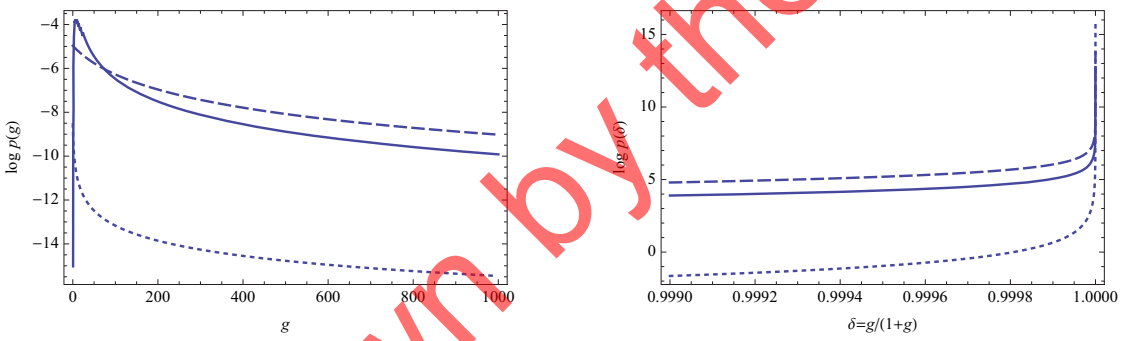
Bottolo-Richardson (solid), Benchmark: $c = 0.1$ (dashed), Hyper-g: $a = 3$ (dotted)



Maruyama-George: $m = 7$ (solid), Feldkircher-Zeigner: $a = 2 + 2/n$ (dashed), Zellner-Siow (dotted)



Forte *et al.*: $m = 7$ (solid), Hyper-g/n: $a = 3$ (dashed), Feldkircher-Zeigner: $a = 2 + 2/k^2$ (dotted)



Hyper-g: $a = 4$ (solid), Horseshoe (dashed), Horseshoe/n (dotted)

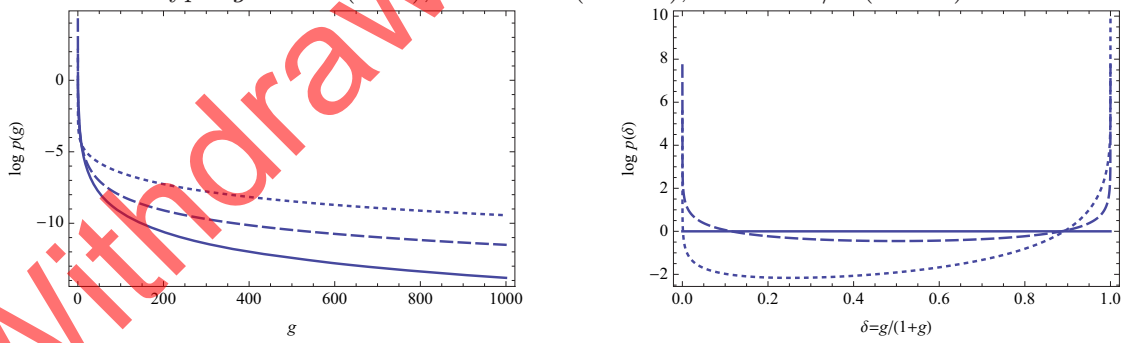


Fig. 2. Priors for g and δ ($n = 72$, $k = 41$)

3.4. Consistency and Information Paradox

As stressed in Liang *et al.* (2008), the limiting behaviour of the Bayes factors can be an interesting guideline for the choice of priors on g. They mention the "information paradox," which occurs if a model M_j gets overwhelming data support (so that $R_j^2 \to 1$) and if then the Bayes factor of M_j with respect to the null model does not go to infinity. It is clear from (4) that this Bayes factor would tend to $\int (1+g)^{(n-1-k_j)/2} p(g) \mathrm{d}g$. Liang *et al.* (2008) show that the Zellner-Siow prior and the hyper-g prior with $a \leq n - k_j + 1$

avoid the paradox, whereas no fixed g would do so. If we adopt the general Beta shrinkage prior, we obtain the result that the paradox is removed if and only if $c < (n-1-k_j)/2$, which exactly corresponds to the findings of Liang *et al.* (2008). Typical values of c chosen in the benchmark Beta prior would certainly comply with this condition.

Consistency implies that all posterior mass tends to be allocated to the true model (*i.e.*, the model that generated the data) if the latter is in the model space as the number of observations goes to infinity. This was introduced in Fernández *et al.* (2001a) and is called "model selection consistency" in Liang *et al.* (2008). The latter paper remarks that the Zellner-Siow and hyper-g/n priors are consistent, but the hyper-g prior is not consistent when the null model is the true model. In contrast, the benchmark Beta prior does lead to consistency in this case:

Proposition 1. *If data are generated by the null model and we adopt the prior for g in (7), the Bayes factor for any other model M_j versus the null model tends to*

$$\frac{\Gamma(b+c)}{\Gamma(b+c+(k_j/2))}\frac{\Gamma(c+(k_j/2))}{\Gamma(c)}$$

as the number of observations n tends to infinity. Thus, if we take the benchmark Beta prior settings in Subsection 3.3 (where $b \to \infty$ with n and c is fixed), we achieve consistency under the null model.

Actually, Proposition 1 can also be used to show that the hyper-g proposal by Feldkircher and Zeugner (2009) where $a = 2 + w(n)$ where $w(n)$ tends to zero with n is consistent, as also mentioned in their paper. With this prior, the Bayes factor for any model versus the null model is $w(n)/(w(n)+k_j)$ and will thus tend to zero.

The Bottolo-Richardson prior has a hyper-g form but depends on n through the truncation and this can be shown to lead to consistency. However, the Bayes factor for any model M_j versus the true null model is or the order $1/\ln(n)$ and the rate of convergence is thus quite slow.

4. Sample information on g and Bayes factors

When we move from choosing a fixed value for g towards the treatment of g as a random quantity with a host of different possible priors, it makes sense to consider what exactly we can expect to learn from the sample about g. Clearly, if g is close to unidentified, we are not going to move away substantially from the prior and the choice of prior is going to be critical. Of course, that does not preclude inference with a proper prior on g but it does mean we should choose the prior carefully.

The information that the sample provides marginally on g can, in principle, be derived from the likelihood integrated with respect to the prior on model parameters in (2):

$$l_y(g, M_j) \propto (1+g)^{\frac{n-1-k_j}{2}}[1+g(1-R_j^2)]^{-\frac{n-1}{2}}, \tag{9}$$

from which we can write the likelihood marginalized with everything except for g as

$$l_y(g) \propto \sum_{j=1}^{2^k}(1+g)^{\frac{n-1-k_j}{2}}[1+g(1-R_j^2)]^{-\frac{n-1}{2}}P(M_j|g). \tag{10}$$

A plot of $l_y(g)$ as a function of g describes the marginal information that is present in the sample about g. With the exception of the priors in Maruyama and George (2010) and Forte *et al.* (2010), the priors on model space do not depend on g and easy analytical expressions for $P(M_j)$ exist, in the case of fixed and random prior covariate inclusion probabilities (Ley and Steel, 2009). The problem in evaluating (10), however, is the huge amount of terms in the sum, which makes this seemingly simple calculation infeasible for $k > 20$ or so.[4]

A feasible way of calculating $l_y(g)$ is to simply start from Bayes rule to realise that

$$l_y(g) = c_i \frac{p_i(g|y)}{p_i(g)}, \tag{11}$$

where we have indicated dependence on the prior used for g by a subscript i. Note that $l_y(g)$ does not depend on the prior chosen for g. For the evaluation of the sample information, the proportionality constant in (11) does not matter, as we are really only interested in the profile of the (unnormalized) integrated likelihood $l_y(g)$ as a function of g.

The posterior density needs to be computed numerically on the basis of an MCMC chain (as described in the next subsection), but this simple equation completely characterizes the information about g contained in the sample. In principle, it should give the same result (up to an additive constant) for any prior on g, but as it is bound to be less precise when both densities tend to zero, an average of (11) will be computed over the different priors used on g, where we discard the influence of a prior for values of g where this prior has very small density values. To neutralise the effect of the different values of c_i in this average, we normalise the profiles by choosing $c_i = p_i(g_0)/p_i(g_0|y)$ for a central value of g_0 for which all priors and posteriors have nonnegligible mass. The more peaked this information measure, the more information the sample contains about g and the less important prior choice on g becomes.

The expression in (11) is also reminiscent of the so-called Savage-Dickey density ratio (Verdinelli and Wasserman, 1995). Indeed, if we make explicit that the proportionality constant $c_i = l_y(\text{prior } i) = \int l_y(g)p_i(g)\text{d}g$, then it becomes clear that the ratio $p_i(g|y)/p_i(g)$ evaluated at a value g_0 is the Bayes factor of the model with fixed g_0 versus the model with prior $p_i(g)$.

Another thing that can be easily done is to compare the data support for different fixed values of g given a choice of prior $p_i(g)$. In particular, if we consider two values, say g_1 and g_2, the Bayes factor is given by

$$\frac{l_y(g_1)}{l_y(g_2)} = \frac{p_i(g_1|y)}{p_i(g_2|y)} \frac{p_i(g_2)}{p_i(g_1)}.$$

Values of g for which the posterior density value is higher relative to the prior density are more strongly supported by the data, in line with intuition. Perhaps even more interesting is the direct comparison of different priors on g, which can be done immediately through the Bayes factors. From (11) the Bayes factor for the model with prior l versus the one with prior m can be computed as

$$\frac{l_y(\text{prior } l)}{l_y(\text{prior } m)} = \frac{p_m(g_0|y)}{p_l(g_0|y)} \frac{p_l(g_0)}{p_m(g_0)}, \tag{12}$$

[4] Note that the sum in (10) is weighted with the prior distribution on model space, so we can not simply approximate it by a sum over the (small subset of) models visited in the MCMC chain used for posterior model inference (see Section 5), but we would typically need complete enumeration.

where we choose a value g_0 such that none of the density values on the right hand side are very small.[5] Note that the data will tend to support models with priors that take relatively high density values at g_0 compared to the posterior. As the marginal likelihood is the likelihood $l_y(g)$ integrated out with the prior $p_i(g)$, it is not surprising that Bayes factors between models that differ in the prior on g will favour those models that concentrate the prior weight on g around values with very high likelihood support.

5. Computational Issues

In typical applications in economics, the number of potential covariates k is so large that the model space is impossible to evaluate exhaustively.[6] So even if Bayes factors between different models can be computed analytically, we still need some sort of numerical method to conduct inference over models. A convenient way to do this is through Markov chain Monte Carlo, in particular the MC^3 algorithm of Madigan and York (1995) has been used in this context with success (Fernández *et al.*, 2001a; and Ley and Steel, 2009). Eicher *et al.* (2010) experiment with various algorithms and find that MC^3 works quite well in this context. Recently, other alternative methods were proposed, such as *Bayesian Adaptive Sampling* by Clyde *et al.* (2010) and *Evolutionary Stochastic Search* by Bottolo and Richardson (2010), which is designed to work for situations where k is orders of magnitude larger than n. We retain the simple MC^3 algorithm, which works well for the types of problems we focus on here.

Except for the prior of Maruyama and George (2010) and Forte *et al.* (2010), we need to deal with the fact that the integral in (4) does not have a straightforward closed-form solution. Liang *et al.* (2008) approximate this integral in g with a Laplace approximation, but we will opt for a Gibbs sampler approach over model space and g. In the latter, the Bayes factor between any two models given g is

$$\frac{l_y(M_i \mid g)}{l_y(M_j \mid g)} = (1+g)^{\frac{k_j-k_i}{2}} \left(\frac{1+g(1-R_i^2)}{1+g(1-R_j^2)} \right)^{-\frac{n-1}{2}}, \tag{13}$$

The conditional posterior of g given M_j is simply[7]

$$p(g \mid y, M_j) \propto (1+g)^{\frac{n-k_j-1}{2}} [1+g(1-R_j^2)]^{-\frac{n-1}{2}} p(g \mid M_j). \tag{14}$$

The advantage of using the Gibbs sampler on (g, M_j) is that it does not rely on approximations that are hard to control and makes prediction quite straightforward: for every g drawn in the sampler we predict as with a fixed g (Fernández *et al.*, 2001a), and predictions are simply mixed over values of g in the sampler. With Laplace approximations this seems much less straightforward and the quality of these approximations is not that easy to assess. Also, truncation, as in the Bottolo-Richardson and Forte *et al.* priors, can be dealt with easily in a Gibbs sampling framework.

[5] The use of the basic identity as in (11) evaluated at fixed parameter values to estimate marginal likelihoods also underlies the proposal of Chib (1995).

[6] For example, the applications to growth data discussed in Section 7 involve model spaces with, respectively, $2^{41} = 2.2 \times 10^{12}$ and $2^{67} = 1.5 \times 10^{20}$ different models.

[7] Indeed, if we take the null model for M_j we get $k_j = 0$ and $R_j^2 = 0$ and we get exactly the prior on g back, as we should (since the null model does not involve g).

We use a simple random walk Metropolis-Hastings step for g with a log-Normal proposal centred over the previous value. Finally, we control the acceptance probability for g by making the Metropolis-Hastings step adaptive.

Throughout, we use MCMC chains of length 1,000,000 after a burn-in of 500,000, which was found to be more than sufficient for convergence. The Fortran code used for this paper is available from the authors upon request.[8]

6. Simulations

First, we examine simulated datasets to mimic the kinds of situations typical for many applications in econometrics (and in other areas as well). We generate datasets from three different model structures, which closely correspond to those in Fernández *et al.* (2001a). For Models 1 and 2 we generate an $n \times k$ matrix R for $k = 15$ regressors where the first ten columns $(r_{(1)}, \ldots, r_{(10)})$ are drawn from independent standard normal distributions, and the next five columns are constructed as

$$(r_{(11)}, \ldots, r_{(15)}) = (r_{(1)}, \ldots, r_{(5)}) \cdot (0.3,\ 0.5,\ 0.7,\ 0.9,\ 1.1)' \cdot \iota_5' + \mathcal{E},$$

where $\mathcal{E}$ denotes an $n \times 5$ matrix drawn from independent normal random variables. This induces correlations between the first five and the last five regressors ranging from 0.153 to 0.74. The demeaned version of R is then the matrix of regressors $Z = (z_{(1)}, \ldots, z_{(15)})$ used to generate the data according to:

$$\text{Model 1:}\ \ y = 4\iota_n + 2z_{(1)} - z_{(5)} + 1.5z_{(7)} + z_{(11)} + 0.5z_{(13)} + \sigma\varepsilon,$$

$$\text{Model 2:}\ \ y = \iota_n + \sigma\varepsilon,$$

where ε is a vector of iid standard normal random variables and we take $\sigma = 2.5$. In Model 1 we use $n = 50, 100$ and for Model 2 we use $n = 50, 100, 1000, 10,000$ in order to illustrate consistency.

The remaining model structure used in the simulations uses regressors with pairwise correlations of 0.5, generated as $r_{(i)} = r^{\star}_{(i)} + e$ where each $r^{\star}_{(i)}$ and e are vectors of n independent standard normal elements. After demeaning to obtain Z we generate n observations as

$$\text{Model 3:}\ \ y = \iota_n + \sum_{h=1}^{7} z_{(h)} + \sigma\varepsilon,$$

where now we use $\sigma = 2$. Values of k used are 20, 40 and 80, while we adopt $n = 50$ and 100.[9]

We analyse 100 simulated data sets for each model and value of k and n mentioned above. The prior on model space is constructed with a random θ as in (5) with a prior mean model size $m = k/2$. The different priors for g discussed above have been used, as well as two fixed values for g ($g = n$ and $g = k^2$). For the hyper-g/n prior we take $a = 3$.

Rather than provide exhaustive results from these simulations, we merely highlight the most important findings here. The different priors do lead to rather different posterior distributions for g. Broadly speaking, the hyper-g priors lead to the smallest median

[8] Fortran code for the fixed g case as described in Ley and Steel (2009) is available online in the *Journal of Applied Econometrics* archive corresponding to the latter paper. Code in R is available from http://bms.zeugner.eu/.

[9] The situation with $n = 50$ and $k = 80$ means that we need to exclude models with $k_j \geq n - 1$, as the posterior for such a model would no longer be well-defined. This is simply done by imposing a prior on model space that limits the model size to $n - 2$. The sampler then rejects any proposed model for which $k_j > n - 2$. In the analysis of these simulated data, the sampler never even gets close to models that large.

posterior median[10] for g, while the benchmark prior with $c = 1$ often leads to the largest. From (13) we can see that if two models, say M_i and M_j, fit equally well, the Bayes factor in favour of M_i given g is approximately $g^{(k_j - k_i)/2}$, so that if M_i has one more regressor than M_j, M_i is penalized by $g^{-1/2}$. Thus a larger g induces a higher model size penalty and this is immediately reflected in the results.

The average number of regressors in the visited models is the first thing we focus on here. Tables 2 and 3 record the median over the 100 samples of the average model size. As expected, average model sizes are smaller for the priors leading to higher values of g, in particular the benchmark prior with $c = 1$ and the case with fixed $g = k^2$. On the other hand, the hyper-g priors and the Bottolo-Richardson and horseshoe priors favour larger models. This is a serious drawback in the case of Model 2, where data are generated from the null model. Here the hyper-g and horseshoe priors (and to a slightly lesser extent the Bottolo-Richardson prior) are choosing models that are far too large: marginal inclusion probabilities of all 15 regressors (none of which are used in the true model) are typically over 0.4 with these priors. In contrast, the probability of erroneously including the regressors tends to be under 1% for the benchmark priors with $c = 1$ and $c = 0.1$ and the case with $g = k^2$. The priors of Maruyama-George, Forte *et al.* and Zellner-Siow and the case with $g = n$ typically lead to marginal inclusion probabilities under 2.5%.

Tables 2 and 3 also present the median posterior probability of the true model. These tend to be smallest for the hyper-g priors throughout and for the Bottolo-Richardson prior in the case of Model 3. The highest probabilities on the true model are typically found for the benchmark prior with $c = 1$ and $c = 0.1$, the Maruyama-George prior and the case with fixed $g = k^2$. For Models 1 and 3 differences in the posterior probability of the true model are not huge (covered by a factor of about 5), but for Model 2 these differences are very substantial indeed, ranging from 0.08 to 0.90 for $n = 50$ and from 0.08 to 0.99 for $n = 10,000$. Table 2 also illustrates the results on consistency, as dealt with in Proposition 1 and summarized in Table 1. The priors not leading to consistency (hyper-g, F-Z with $a = a + 2/k^2$ and horseshoe priors and the case with $g = k^2$) lead to posterior probabilities on the true model that are virtually unaffected by sample size n. The consistent priors (benchmark, Maruyama-George, Bottolo-Richardson, F-Z with $a = 2 + 2/n$, Forte *et al.*, hyper-g/n, horseshoe/n, Zellner-Siow and the case with $g = n$) see this probability increase to unity with n, although this convergence seems relatively slow for the horseshoe/n and very slow for the Bottolo-Richardson prior and the F-Z prior with $a = 2 + 2/n$.[11]

Another consequence of different model size penalties is the number of models that are actually visited by the chain. If we make the prior on model space adaptive to the data by using a hierarchical prior on θ, the model size penalty mainly stems from g (Ley and Steel, 2009). Thus, we would expect the number of models visited by the chain to be affected by the choice of prior on g. Indeed, that is the case, with the hyper-g and horseshoe priors typically leading to large number of models visited and the benchmark prior with $c = 1$ and the case with $g = k^2$ often resulting in much less model visits. The differences can be quite substantial—*e.g.*, in Model 1 with $n = 50$ the hyper-g and horseshoe priors lead to a median of over 10,000 model visits whereas the chains for the benchmark prior with

[10] Here and in the subsequent discussion we often focus on the median over the 100 samples we generated. We do not consider the median of the posterior means of g, since the posterior mean of g does not exist for all but the Bottolo-Richardson prior (see the discussion in Subsection 3.3).

[11] In the case of the latter prior, Proposition 1 tells us that the Bayes factor in favour of the (true) null model against any other model M_j is $1 + nk_j/2$, which seems to be sizeable for $n = 10,000$, but we need to keep in mind that there are many alternative models.

$c = 1$ visit a median of 4128 different models, and if we fix $g = k^2$ this is only 2132 models. For Model 2 we observe even larger differences, with the number of models visited ranging from around 200 for the benchmark prior with $c = 1$, Maruyama-George, Forte *et al.* and $g = n$ to about 33,000 for the hyper-g and the horseshoe priors if we take $n = 1000$. As expected, for Model 3 with $k = 80$ we observe most visits in the chains, ranging from a median visit count of 3280 for the case with $g = k^2$ to around 60,000 for the hyper-g priors.

An interesting aspect that is suggested by Table 3 is the effect of different values of k and n. One would expect that the challenge of finding the model that generated the data is harder as the model space (and thus k) grows. Indeed, if we have only $n = 50$ observations, the posterior probability assigned to the true model decreases by around an order of magnitude each time we double the value of k. However, if we have 100 observations to conduct inference with, we see that the posterior probability of the true model is almost unaffected by the changes in k. There is a slight downward tendency with k, but the median is never even halved by going from $k = 20$ to $k = 80$. This seems to suggest there is a critical level of n, which allows us to identify a relatively small area in model space and above which we are not too much led astray by the introduction of further unrelated regressors. Interestingly, the marginal posterior inclusion probabilities of the irrelevant regressors is not much smaller for the case with $n = 100$, but clearly the entire posterior distribution on model space is quite different.

Table 2. Simulated Data, Models 1 and 2 (null model)—results for model size and inclusion probability of the true model. Number of regressors $k = 15$.

		Model 1				Model 2							
True model size		5		5		0		0		0		0	
Prior mean model size		7.5		7.5		7.5		7.5		7.5		7.5	
n		50		100		50		100		1000		10,000	
		median average posterior model size: η median posterior prob. true model (in %): γ											
	Prior on g	η	γ	η	γ	η	γ	η	γ	η	γ	η	γ
1	Benchmark $c = 1$	4.2	0.4	5.2	3.1	0.1	88	0.1	87	0.1	94	0.0	98
2	Benchmark $c = 0.1$	5.3	0.5	5.9	2.5	0.1	90	0.1	89	0.0	95	0.0	99
3	Benchmark $c = 0.01$	6.2	0.5	6.1	2.3	1.1	56	0.8	45	0.2	89	0.0	97
4	Maruyama-George	5.5	0.5	5.7	2.5	0.3	81	0.1	89	0.0	97	0.0	99
5	Bottolo-Richardson	6.4	0.5	6.2	2.2	4.8	18	4.2	20	3.9	28	3.2	31
6	Hyper-g, $a = 3$	7.4	0.3	6.9	1.7	6.6	9	6.4	9	6.4	9	6.3	9
7	Hyper-g, $a = 4$	7.7	0.3	7.2	1.5	6.8	8	6.6	8	6.7	8	6.6	8
8	F-Z, $a = 2 + 2/k^2$	6.3	0.4	6.1	2.2	2.2	41	2.0	40	2.0	40	2.2	44
9	F-Z, $a = 2 + 2/n$	6.3	0.4	6.2	2.3	2.5	40	2.1	37	2.0	42	2.1	43
10	Horseshoe	6.8	0.4	6.5	2.1	6.5	10	6.2	11	6.3	11	6.3	11
11	Forte *et al.*	6.0	0.5	6.0	2.4	0.3	78	0.2	86	0.0	96	0.0	99
12	Hyper-g/n	5.8	0.5	5.8	2.4	1.2	44	0.7	53	0.1	91	0.0	98
13	Horseshoe/n	6.1	0.4	6.0	2.4	3.9	29	3.1	35	1.7	54	0.6	66
14	Zellner-Siow	5.6	0.5	5.9	2.4	0.4	73	0.2	82	0.1	94	0.0	98
15	$g = n$	4.8	0.6	5.1	3.1	0.4	71	0.2	80	0.1	94	0.0	98
16	$g = k^2$	3.6	0.3	4.6	3.6	0.2	87	0.2	86	0.2	86	0.1	87

Combining the results on consistency with those of the posterior probability assigned to the true model, we conclude from this section that the best performing priors seem to be the benchmark, Maruyama-George, Forte *et al.* and the Zellner-Siow priors.

Table 3. Simulated Data, Model 3—results for model size and inclusion probability of the true model.

		Model 3											
		$k = 20$				$k = 40$				$k = 80$			
True model size		7		7		7		7		7		7	
Prior mean model size		10		10		20		20		40		40	
n		50		100		50		100		50		100	
		median average posterior model size: η median posterior prob. true model (in %): γ											
	Prior on g	η	γ	η	γ	η	γ	η	γ	η	γ	η	γ
1	Benchmark $c = 1$	6.7	1.1	8.1	23	5.3	0.1	7.6	29	4.1	0.00	7.1	30
2	Benchmark $c = 0.1$	7.0	1.0	8.2	21	6.3	0.1	8.1	20	5.3	0.01	7.9	19
3	Benchmark $c = 0.01$	7.1	1.0	8.2	20	6.6	0.1	8.2	18	6.1	0.01	8.3	15
4	Maruyama-George	7.1	1.0	8.2	20	6.7	0.1	8.3	18	6.2	0.01	8.4	14
5	Bottolo-Richardson	7.9	0.5	9.1	10	7.3	0.1	9.0	9	6.7	0.00	9.2	6
6	Hyper-g, $a = 3$	7.5	0.8	8.5	16	7.0	0.1	8.5	15	6.5	0.00	8.6	11
7	Hyper-g, $a = 4$	7.6	0.8	8.6	15	7.1	0.1	8.6	14	6.6	0.01	8.7	10
8	F-Z, $a = 2 + 2/k^2$	7.1	1.0	8.2	20	6.7	0.1	8.3	18	6.2	0.01	8.4	13
9	F-Z, $a = 2 + 2/n$	7.1	1.0	8.2	20	6.7	0.1	8.3	18	6.2	0.01	8.3	14
10	Horseshoe	7.3	1.0	8.3	18	6.8	0.1	8.3	17	6.4	0.01	8.5	12
11	Forte *et al.*	7.3	0.9	8.3	18	6.8	0.1	8.4	17	6.4	0.01	8.5	12
12	Hyper-g/n	7.1	1.1	8.2	20	6.7	0.1	8.3	18	6.1	0.01	8.4	13
13	Horseshoe/n	7.2	1.0	8.2	19	6.7	0.1	8.3	18	6.3	0.01	8.4	13
14	Zellner-Siow	7.2	1.0	8.3	19	6.8	0.1	8.3	18	6.3	0.01	8.4	13
15	$g = n$	7.5	0.8	8.7	14	7.1	0.1	8.6	13	6.6	0.01	8.8	8
16	$g = k^2$	6.2	0.9	7.8	31	4.9	0.0	7.2	36	3.8	0.00	6.6	28

7. Applications to Real Data

In the context of real applications, we will further examine all the different priors, and will also investigate the predictive performance as well as Bayes factors. For each application, we will randomly partition the sample in an estimation sample and a prediction sample, of a fixed size. We do this 50 times and this allows us to assess predictive performance based on the log predictive score, abbreviated to LPS (Fernández *et al.*, 2001a,b).[12] In addition, analyzing the 50 estimation subsamples allows us to get a certain amount of robustness with respect to possible unusual data points, partly addressing some of the concerns voiced in Ciccone and Jarociński (2010)[13] with respect to data sensitivity. Bayes factors between models with different priors will be computed on the basis of the full sample. We consider four different priors on the model space: fixed θ and random θ as in (5), both with prior mean model sizes of $m = 7$ and $m = k/2$.

7.1. Cross-country growth regressions

The context of cross-country growth regressions is one of the first areas within economics where the use of BMA has become popular. This area of macroeconomics is characterized by a particularly large number of potential drivers for growth and a scarcity of observations,

[12] Alternative proper scoring rules are discussed in Gneiting and Raftery (2007).

[13] However, also see the comment in Feldkircher and Zeugner (2010).

so this is an ideal candidate for BMA. We consider below two datasets which have been used in many studies on this topic.

The FLS data

The first dataset we use contains $k = 41$ potential regressors for modelling average per capita growth over the period 1960-1992 for a sample of $n = 72$ countries. It was used in Fernández *et al.* (2001b) (FLS), which presents more details.

First, we consider the analysis of the 50 subsamples which each consist of 62 observations. In line with the higher model size penalty implicit in the benchmark Beta prior with $c = 1$ and the case with fixed $g = k^2$, we note that marginal posterior inclusion probabilities of the regressors tend to be (sometimes much) smaller for these cases. In keeping with the higher g values induced by these priors, the posterior mean of β_j for model M_j will be larger in absolute value (less shrinkage to the prior mean of zero), but this will typically be counteracted by more mass on smaller models. The resulting marginal posterior means of the regression coefficients can be smaller or larger than those for the other priors on g. For the other priors posterior inclusion probabilities and posterior inference on the regression coefficients are quite similar, even though the posterior inference on g can be somewhat different. Figure 3 presents boxplots over the 50 subsamples for the posterior medians of g (on a log scale), the number of visited models and the posterior mean model size. As with the simulated data, posterior medians of g are smallest for the hyper-g priors and largest for the benchmark prior with $c = 1$ and the case with fixed $g = k^2$. This leads to most models visited in the chain for the hyper-g priors and least models in the chain for the benchmark with $c = 1$ and $g = k^2$ cases. Posterior model sizes are lower for the latter two cases, but not that different between the other priors. Note also that the prior assumptions on model space have almost no effect on posterior inference on g, but prior mean model size m does substantially affect the number of visited models and the posterior mean model size when we use a fixed θ approach. In line with the results in Ley and Steel (2009), however, this dependence on m virtually disappears when we use the random θ prior in (5). Thus, for the rest of the section we will only present results for random θ with $m = 7$.

Prediction based on LPS (for 10 observations in each prediction subsample) is summarized in the top panel of Figure 4, where we present boxplots of the LPS values for BMA over the 50 prediction subsamples. As lower values of LPS are associated with better predictions, this suggests that the benchmark prior with $c = 1$ and the fixed $g = k^2$ case tend to predict somewhat worse than the rest. The other priors are quite close in predictive performance. This is consistent with these priors leading to quite similar posterior inclusion probabilities for the regression variables, and rather different from those obtained with the priors that result in worse predictions. BMA predicts best (compared with the highest posterior probability model, the null model and the full model) in most of the subsamples (in between 52% and 62%, depending on which prior we choose for g) and is never beaten by the null model, except for a few cases (up to 6%) with the benchmark priors with $c = 1$ and $c = 0.1$ and the fixed g specifications.

The information contained in the data regarding g is summarized by the marginal likelihood $l_y(g)$ in (10). Figure 5 plots an estimate computed as in Section 4. Clearly, the data do possess some information on g, and particularly favour values in the region 15 to 50.

Bayes factors of the models with different priors over g can be computed as described in (12) in Section 4. The particular value g_0 at which we evaluate this expression is not very critical to the outcome. Here we use the mode of $l_y(g)$ as a reasonable value for g_0.

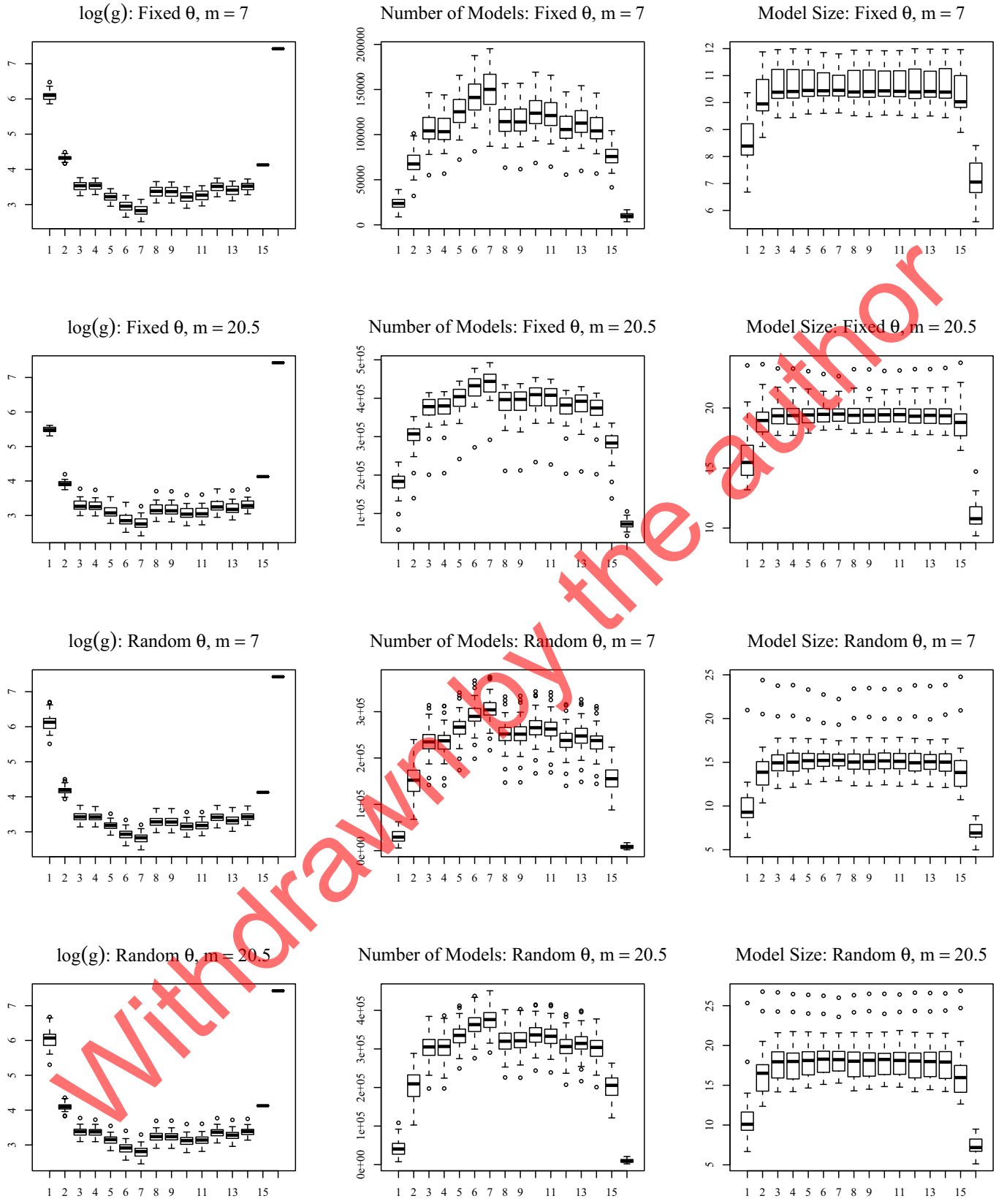


Fig. 3. FLS data: log posterior medians of g, number of models visited and posterior mean model size
The different priors on g are ordered as in Tables 2 and 3. The top two rows correspond to fixed θ (with different m), while the bottom two rows are for random θ priors on model space

Table 4 lists the Bayes factors for the models with different priors for g versus the one with

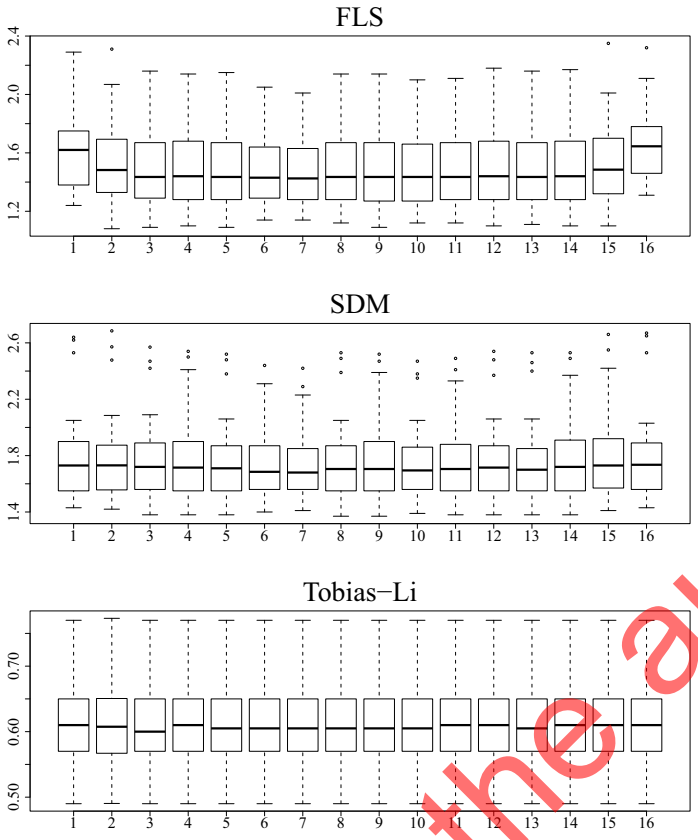


Fig. 4. Log Predictive Score: FLS, SDM and Tobias-Li data (random θ, $m = 7$)

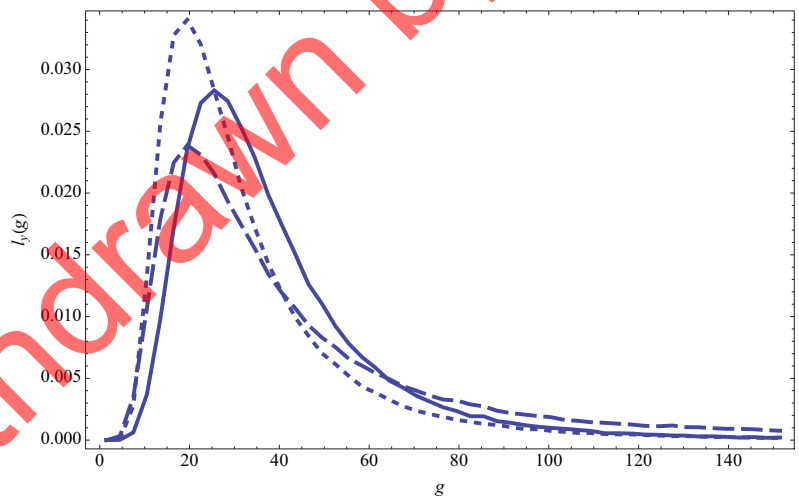


Fig. 5. Marginal likelihood $l_y(g)$: FLS (solid), SDM (dashed) and Tobias-Li (dotted) data (random θ, $m = 7$)

the benchmark Beta prior with $c = 0.01$ (Prior 3), computed over the full sample. The benchmark prior with $c = 1$ (Prior 1) leads to very small prior and posterior density values at g_0, so that the ratio can really not be computed with any accuracy. However, given that this is indicative of a very small prior density in the region of the mode of $l_y(g)$, it is safe to assume the associated Bayes factor would be quite small. The same problem affects other models with the other two datasets (indicated with – in Table 4). As commented in Section 4, the priors that put a lot of weight on regions of g with high $l_y(g)$ do particularly well

in this respect. Thus, the Forte *et al.*, Zellner-Siow, Bottolo-Richardson, hyper-g/n and horseshoe/n priors lead to the highest Bayes factors while the F-Z, hyper-g and benchmark priors do worse. This is in line with the prior density functions in Figures 1 and 2 and the marginal likelihood in Figure 5. In Table 4 we also include the models where we choose a fixed value of g, in which case we can compute the Bayes factor as the ratio $p_3(g \mid y)/p_3(g)$ evaluated at the fixed value of g (as explained in Section 4). Some entries in the table are missing, as they correspond to $p_3(g|y) = 0$ for the fixed value of g; again, this can be taken as a sign that this value is associated with a very small Bayes factor. So, whereas the data do support choosing $g = n$, the case $g = k^2$ does not receive much support in terms of Bayes factors.

Table 4. Bayes Factors of Different Priors vs Benchmark with $c = 0.01$ (Prior 3) (Random θ, $m = 7$)

i	Prior on g	FLS	SDM	Tobias-Li
2	Benchmark $c = 0.1$	0.3	–	0.4
4	Maruyama-George	16.3	10.1	–
5	Bottolo-Richardson	23.6	57.2	24.4
6	Hyper-g, $a = 3$	2.3	8.7	3.0
7	Hyper-g, $a = 4$	0.7	3.0	1.0
8	F-Z, $a = 2 + 2/k^2$	0.1	0.1	0.3
9	F-Z, $a = 2 + 2/n$	2.4	5.6	0.1
10	Horseshoe	11.2	34.6	11.9
11	Forte *et al.*	53.2	123.2	n.a.
12	Hyper-g/n	21.8	50.8	1.7
13	Horseshoe/n	25.9	68.5	7.6
14	Zellner-Siow	29.8	60.1	–
	g_0	25.5	19.5	19.5
15	$g = n$	19.7	37.1	–
16	$g = k^2$	–	0.1	–

The SDM data

Another popular growth dataset was introduced by Sala-i-Martin *et al.* (2004) (SDM), who model annual GDP growth per capita between 1960 and 1996 for $n = 88$ countries using $k = 67$ potential regressors.

The 50 randomly selected estimation subsamples each have 75 observations, and lead to the boxplots in Figure 6. The results are mostly in line with those for the FLS data, but with larger differences between the priors in terms of posterior model size.

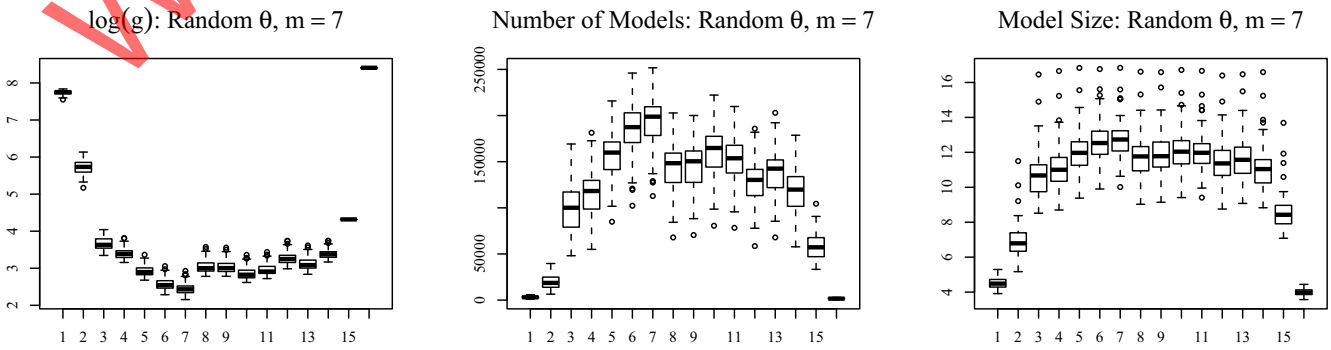


Fig. 6. SDM data: log posterior medians of g, number of models visited and posterior mean model size.

Predictions for these data seem to be quite similar for all priors, as judged by the LPS boxplots in Figure 4. Here we have used prediction subsets of 13 observations each. Interestingly, these comparable prediction accuracies are obtained with regression models that are rather different, since the posterior inclusion probabilities corresponding to the benchmark priors with $c = 1$ and $c = 0.1$ and to the choice of $g = k^2$ (*i.e.*, the models leading to the highest values for g) are quite different from those using the other priors. BMA predicts best in 76% to 92% of the subsamples, with the exception of the benchmark priors with $c = 1$ and $c = 0.1$ and the case with $g = k^2$, where BMA outperforms the highest posterior probability model, the null model and the full model in roughly half the subsamples. BMA is only beaten by the null model in up to 8% of the cases,[14] except for the benchmark priors with $c = 1$ and $c = 0.1$ and the fixed-g specifications, where the null model outperforms BMA in 10% or 12% of the cases.

The marginal information in the data on g, as presented in Figure 5, is quite similar to that for the LPS data, with a slight shift to smaller values and a heavier tail. As a consequence, the pattern of Bayes factors between the models with different priors on g is rather similar to what they were for the FLS data.

7.2. Returns to Schooling

Here we investigate the area of returns to education, where again many potential model specifications have been proposed, and BMA was introduced in this context by Tobias and Li (2004). As these are microeconomic data, the number of potential observations is much larger. We will focus on the log of hourly wages of $n = 1190$ white males in the U.S. in 1990, which we model as a function of $k = 26$ possible regressors. In order to simplify the computations, Tobias and Li (2004) always include seven of these regressors in each model and restrict themselves to a smaller model space characterized by 18 possibly included regressors, but we allow for inclusion or exclusion of all regressors. In addition, we have added the local unemployment rate (which only appears in Tobias and Li, 2004, through an interaction term with education) as a potential regressor, giving us a total of 26 candidate regressors. More details on the application can be found in Tobias and Li (2004).

On the basis of 50 subsamples of 1012 observations, we get the results for g, numbers of models visited and model size presented in Figure 7. Results are now a bit different from those in the growth applications, as a result of the much larger value of n. In particular, the Maruyama-George, Forte *et al.* and Zellner-Siow priors as well as the case with fixed $g = n$ lead to much higher posterior medians of g, a smaller number of models visited in the chain and smaller mean model size than in the previous applications. Of course, all of these priors depend on n, but it is interesting that the F-Z prior with $a = 2 + 2/n$ and the hyper-g/n and horseshoe/n priors are less affected. Of course, the benchmark priors and the Bottolo-Richardson prior depend on n through $\max\{n, k^2\}$, which is here of the same order of magnitude as for the FLS data.

Prediction based on prediction subsamples of 178 observations each is summarized in the lower panel of Figure 4. Even though there is some variability in the posterior inclusion probabilities of the regressors over the different priors, we observe virtually identical predictive performance. This is related to the fact that the full model here predicts very similarly to using BMA. This is not surprising as the number of observations n is much larger than the number of potential regressors, so we can hardly overfit and for prediction

[14] For the hyper-g prior with $a = 4$, BMA is actually never beaten by the null model.

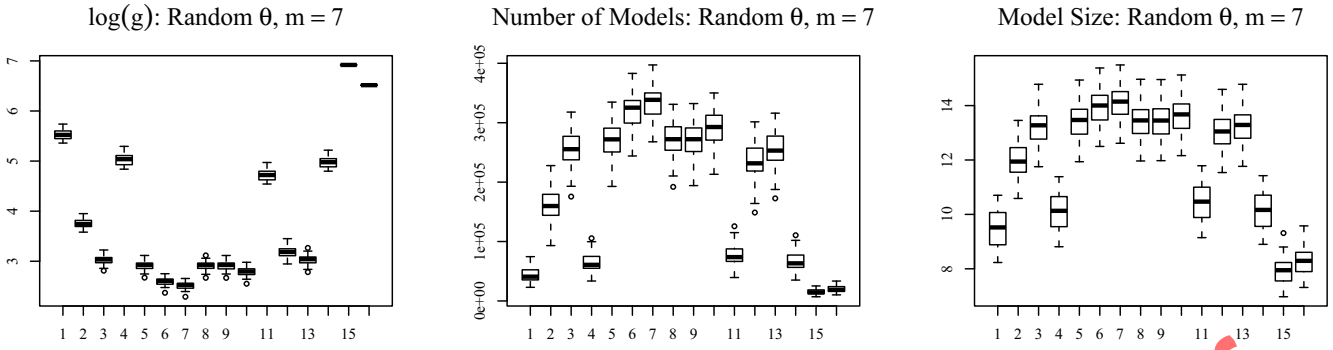


Fig. 7. Tobias-Li data: log posterior medians of g, number of models visited and posterior mean model size.

not much is lost by including all k regressors. In the full model, the only effect that the prior for g has on prediction is the different shrinkage of the coefficients, but as the shrinkage coefficient is very close to one anyway, that will make very little difference. Of course, BMA still has an important role in identifying the important determinants of wages, but is not that critical for prediction. In line with this, the null model does very badly here and never beats BMA or the full model. BMA predicts best in 20% to 46% of the subsamples, with the lowest percentages occurring for the benchmark Beta priors with $c = 1$ and the fixed-g cases. BMA is beaten by the full model in 42% to 48% of the cases with random g. For the two specifications with fixed g, the full model beats BMA in 52% of the subsamples, while the best model also beats BMA there in over a third of the cases.

Figure 5 tells us that the likelihood marginalized with the prior of everything but g is rather similar to that of the growth datasets. This is perhaps surprising, given that the number of observations n is quite a lot larger in this case. This leads to somewhat similar behaviour of the Bayes factors as for the two growth datasets, with the exception of the Maruyama-George and Zellner-Siow priors (leading to a zero posterior density at g_0), the F-Z with $a = 2 + 2/n$, the hyper-g/n and horseshoe-n priors and the case with $g = n$, as a consequence of the very different value of n for this application. In this application, the truncation induced by the Forte *et al.* prior means that the prior support does not include the chosen value of g_0 (and is truncated quite a bit above that), which certainly makes this prior less appropriate for large values of n/k. Thus, prior and posterior density values are both zero at g_0 and the Bayes factor can not be computed for this case.[15]

Let us summarize the information that these applications to a variety of real economic datasets are providing us with. Predictive performance seems quite similar across priors, although possibly a bit worse for the benchmark prior with $c = 1$ and the case with $g = k^2$. There is more variation in the Bayes factors, with most consistent data support provided for the Bottolo-Richardson, horseshoe, hyper-g/n and horseshoe/n priors, and the Maruyama-George, Forte *et al.* and Zellner-Siow priors (as well as the case with fixed $g = n$) doing quite well in two of the three applications, but rather badly in the one with large n.

[15] Using a much larger value of g_0 ($g_0 > 40$) it would be possible, in principle, to compute a Bayes factor, but we can safely assume the Tobias-Li data do not support this prior at all, as it puts no mass whatsoever on areas with large values for the marginal likelihood $l_y(g)$.

8. Concluding Remarks and Recommendations

Combining the properties listed in Table 1 with the evidence from both simulated and real data, we can now come up with a recommendation for practitioners. We assume that users will want their priors to be consistent, to avoid the information paradox and to perform well in a wide variety of situations. The Bottolo-Richardson and horseshoe-n priors are consistent (albeit at a slow rate) and do well on real data, but underperform on the simulated data. The truncation of the Forte *et al.*. prior makes it hard to recommend for situations where n is appreciably larger than k. The benchmark Beta prior with $c = 1$ does not fare well in prediction for one application, and the benchmark prior with $c = 1$ and $c = 0.1$ does not do well in terms of Bayes factors. The Zellner-Siow and Maruyama-George priors do quite well, except for the last application (with large n), where they get very little support from the data. Nevertheless, we feel they do deserve a place in the econometrician's toolbox, especially if n is relatively small (comparable to k, say). In addition, if one wants to avoid numerical methods to compute Bayes factors between models with different sets of regressors, the Maruyama-George prior should be recommended, although this comes at the (small) cost of making the prior model-specific, which may make it slightly harder to interpret the prior on g.

In our view, the two priors that stand out by not having displayed any truly bad behaviour in our experiments are the benchmark Beta prior with $c = 0.01$ and the hyper-g/n prior (with $a = 3$).[16] Thus, these priors provide an interesting compromise and would be our general recommendations to practitioners.

The hierarchical Bayesian model explored in this paper has a hyperprior on prior variable inclusion probabilities and a hyperprior on g, which leads to an integral for the marginal likelihood that is solved by running the MCMC sampler over models and g jointly. The advantage of these hyperpriors is to make the analysis more robust with respect to often arbitrary prior assumptions. We now allow the data to inform us on variable inclusion probabilities and the appropriate region for g. The latter will have an effect on both the model size penalty and on shrinkage for each given model. Putting a prior on g makes these things naturally adaptive and avoids the information paradox (Liang *et al.*, 2008), which affects analyses with fixed g. We feel the model used here with the recommended priors on g can be considered as a safe "automatic" prior for use in Bayesian Model Averaging in the types of linear regression problems that typically arise in a variety of econometric settings.

[16] For the hyper-g prior we find (as do Feldkircher and Zeugner, 2009) that the choice between $a = 3$ and $a = 4$ makes little difference. As a consequence, we would also expect the hyper-g/n prior with $a = 4$ to perform very similarly.

References

Andrews, D.R., and C.L. Mallows (1974) "Scale Mixtures of Normal Distributions," *Journal of the Royal Statistical Society*, B, 36: 99–102.

Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., New York: Springer.

Bernardo, J.M., and A.F.M. Smith (1994) *Bayesian Theory*, Chichester: John Wiley.

Bottolo L., and Richardson S. (2008), "Fully Bayesian Variable Selection Using g-Priors," Working paper, Imperial College.

Bottolo L., and Richardson S. (2010), "Evolutionary Stochastic Search for Bayesian Model Exploration," *Bayesian Analysis*, 5: 583–618.

Brock, W., S. Durlauf and K. West (2003) "Policy Evaluation in Uncertain Economic Environments," (with discussion) *Brookings Papers of Economic Activity*, 1: 235–322.

Brown, P.J., M. Vannucci and T. Fearn (1998) "Bayesian Wavelength Selection in Multicomponent Analysis," *Journal of Chemometrics*, 12: 173–182.

Carvalho, C.M., N.G. Polson and J.G. Scott (2010) "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97: 465–480.

Chib, S. (1995) "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90: 1313–1321.

Ciccone, A. and Jarociński, M. (2010). Determinants of Economic Growth: Will Data Tell? *American Economic Journal: Macroeconomics*, 2: 222–246.

Clyde, M.A., and E.I. George (2004) "Model Uncertainty," *Statistical Science*, 19: 81–94.

Clyde, M.A., J. Ghosh and M. Littman (2010) "Bayesian adaptive sampling for variable selection and model averaging", *Journal of Computational and Graphical Statistics*, forthcoming.

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

Eicher, T.S., C. Papageorgiou and A.E. Raftery (2010) "Determining Growth Determinants: Default Priors and Predictive Performance in Bayesian Model Averaging, With Applications to Growth Determinants," *Journal of Applied Econometrics*, forthcoming.

Feldkircher, M. and S. Zeugner (2009) "Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging," IMF Working Paper 09/202.

Feldkircher, M. and S. Zeugner (2010) "The Impact of Data Revisions on the Robustness of Growth Determinants: A Note on 'Determinants of Economic Growth. Will Data Tell'?," Working Paper 2010-12, University of Salzburg.

Fernández, C., E. Ley and M.F.J. Steel (2001a) "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100: 381–427.

Fernández, C., E. Ley and M.F.J. Steel (2001b) "Model Uncertainty in Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16: 563–76.

Fernández, C., and M.F.J. Steel (2000) "Bayesian Regression Analysis With Scale Mixtures of Normals," *Econometric Theory*, 16: 80-101

Foster, D.P., and E.I. George (1994), "The Risk Inflation Criterion for multiple regression," *Annals of Statistics*, 22: 1947–1975.

Forte, A., M.J. Bayarri, J.O. Berger and G. García-Donato (2010), "Closed-form objective Bayes factors for variable selection in linear models", poster presentation *Frontiers of Statistical Decision Making and Bayesian Analysis* in honour of Jim Berger.

Gneiting, T. and A.E. Raftery (2007) "Strictly Proper Scoring Rules, Prediction and Estimation," *Journal of the American Statistical Association*, 102: 359–378.

Gradshteyn, I.S. and I.M. Ryzhik (1994) *Table of Integrals, Series and Products*, San Diego: Academic Press, 5th ed.

Hoeting, J.A., D. Madigan, A.E. Raftery and C.T. Volinsky (1999) "Bayesian model averaging: A tutorial," *Statistical Science* 14: 382–401.

Jeffreys, H. (1961) *Theory of Probability*, Oxford: Clarendon Press, 3rd ed.

Kass, R.E. and L. Wasserman (1995) "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," *Journal of the American Statistical Association*, 90: 928–934.

Liang, F., R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger (2008) "Mixtures of g-priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103: 410–423.

Ley, E. and M.F.J. Steel (2009) "On the Effect of Prior Assumptions in Bayesian Model Averaging With Applications to Growth Regression," *Journal of Applied Econometrics*, 24: 651–674.

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

Maruyama, Y. and E.I. George (2010) "gBF: A Fully Bayes Factor with a Generalized g-prior," Technical Report, University of Pennsylvania, available at `http://arxiv.org/abs/0801.4410`

Mitchell, T.J. and J.J. Beauchamp (1988) "Bayesian variable selection in linear regression (with discussion), " *Journal of the American Statistical Association*, 83: 1023–1036.

Nott, D.J. and R. Kohn (2005) "Adaptive Sampling for Bayesian Variable Selection," *Biometrika*, 92: 747–763

Raftery, A.E., D. Madigan, and J. A. Hoeting (1997) "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92: 179–191.

Sala-i-Martin, X.X., G. Doppelhofer and R.I. Miller (2004) "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94: 813–835.

Strawderman, W. (1971) "Proper Bayes minimax estimators of the multivariate normal mean," *Annals of Statistics*, 42: 385–388.

Tobias, J.L. and Li, M. (2004) "Returns to Schooling and Bayesian Model Averaging; A Union of Two Literatures," *Journal of Economic Surveys*, 18: 153–180.

Verdinelli, I. and Wasserman, L. (1995) "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, 90: 614-618.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.

Zellner, A. (1986) "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, eds. P.K. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233–243.

Zellner, A. and Siow, A. (1980) "Posterior odds ratios for selected regression hypotheses," (with discussion) in *Bayesian Statistics*, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Valencia: University Press, pp. 585–603.