TEXed on January 7, 2008 at 7:20 A.M. by E. LEY

On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression

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

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

Version: January 7, 2008

Abstract. We consider the problem of variable selection in linear regression models. Bayesian model averaging has become an important tool in empirical settings with large numbers of potential regressors and relatively limited numbers of observations. We examine the effect of a variety of prior assumptions on the inference concerning model size, posterior inclusion probabilities of regressors and on predictive performance. We illustrate these issues in the context of cross-country growth regressions using three datasets with 41 to 67 potential drivers of growth and 72 to 93 observations. Finally, we recommend priors for use in this and related contexts.

Keywords. Model size; Model uncertainty; Posterior odds; Prediction; Prior odds; Robustness

JEL Classification System. C11, O47

Address. E. Ley, The World Bank, Washington DC 20433. Email: eley@worldbank.org. URL: http://eduley.org/. M.F.J. Steel, Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K. Email: M.F.Steel@stats.warwick.ac.uk Vox: +44(0)24 7652 3369. Fax: +44(0)24 7652 4532. URL: http://www.warwick.ac.uk/go/msteel/

1. Introduction

This paper considers model uncertainty associated with variable selection in linear regression models. In particular, we focus on applications to cross-country growth regressions, where we often face a large number of potential drivers of growth with only a limited number of observations. Insightful discussions of model uncertainty in growth regressions can be found in Brock and Durlauf (2001) and Brock, Durlauf and West (2003). Various approaches to deal with this model uncertainty have appeared in the literature, starting with the extreme-bounds analysis in Levine and Renelt (1992) and the confidence-based analysis in Sala-i-Martin (1997). A natural solution, supported by formal probabilistic reasoning, is the use of Bayesian model averaging (BMA, see Hoeting *et al.*, 1999), which assigns probabilities on the model space and deals with model uncertainty by mixing over models, using the posterior model probabilities as weights.

Fernández *et al.* (2001b, FLS henceforth) introduce the use of BMA in growth regressions. In this context, the posterior probability is often spread widely among many models, which strongly suggests using BMA rather than choosing a single model. Evidence of superior predictive performance of BMA can be found in, *e.g.*, Raftery *et al.* (1997), Fernández *et al.* (2001a) and FLS. Other papers using BMA in the context of growth regression are León-González and Montolio (2004) and Masanjala and Papageorgiou (2005, MP henceforth). Alternative ways of dealing with model uncertainty are proposed in Sala-i-Martin *et al.* (2004, SDM henceforth), and Tsangarides (2005). As we will show in the paper, the SDM approach corresponds quite closely to a BMA analysis with a particular choice of prior.

For Bayesian (and approximately Bayesian) approaches to the problem, any differences can typically be interpreted as the use of different prior assumptions. A casual comparison of results can sometimes lead to a misleading sense of "robustness" with respect to such assumptions. In particular, posterior results on inclusion probabilities of regressors reported in SDM were found to be rather close to those obtained with the quite different prior settings of FLS, using the same data; such similarities were noted in MP and Ley and Steel (2007). As we will show here, this is mostly by accident, and prior assumptions can be extremely critical for the outcome of BMA analyses—this general issue is also studied in Eicher *et al.* (2007), which came to our attention after writing the present paper. As BMA or similar approaches are rapidly becoming mainstream tools in this area, we wish to investigate in detail how the (often almost arbitrarily chosen) prior assumptions may affect our inference. We need to stress at this stage that the prior structures examined in this paper are limited to the most commonly used "vague" choices, that require only a minimal amount of prior elicitation and do not cover more informative prior structures, such as the hierarchical model prior structures of Brock *et al.* (2003).

As a general principle, the effect of not strongly held prior opinions should be minimal. This intuitive sense of a "non-informative" or "ignorance" prior is often hard to achieve, especially when we are dealing with model choice, as opposed to inference within a given model—see, *e.g.*, Kass and Raftery (1995). At the very least, we should be able to trace the effect of these assumptions in order to inform the analyst which prior settings are more informative than others, and in which

We are grateful for useful comments from Steven Durlauf, Theo Eicher, Chris Papageorgiou, Adrian Raftery, Chris Sims, two anonymous referees, and participants in seminars at the University of Washington, Princeton University, the Sixth Workshop on Objective Bayes Methodology in Rome, the Summer 2007 European Econometric Society meetings in Budapest, and The World Bank.

direction they will influence the result. “Clever” prior structures are robust, in that they protect the analyst against unintended consequences of prior choices. In this paper, we focus on a general prior structure which encompasses most priors used in the growth regression literature and allows for prior choice in two areas: the choice of the precision factor g in the g-prior, and the prior assumptions on the model space. On the latter, we elicit the prior in terms of the prior mean model size m, which is a quantity that analysts may have some subjective prior preferences on. Other aspects of the prior are typically less interpretable for most applied analysts and would require “automatic” settings. However, these choices have to be reasonable and robust. It is important to stress that the dependence on prior assumptions does not disappear if we make those assumptions implicit rather than explicit. We then merely lull the analyst into a false sense of security. Thus, the claim in SDM that their approach “limits the effect of prior information” has to be taken with extreme caution.

To build priors on the model space, we shall advocate the use of hierarchical priors, since this increases flexibility and decreases the dependence on essentially arbitrary prior assumptions. Theoretical results on the distribution of model size and prior odds allow us to shed some light on the relative merits of the priors on model space. Analytical results for the marginal likelihoods (Bayes factors) are used to infer the model size penalties implicit in the various choices of g and allow for an insightful comparison with the BACE procedure of SDM.

Using three different data sets that have been used in the growth literature, we assess the effect of prior settings for posterior inference on model size, but we also consider the spread of model probabilities over the model space, and the posterior inclusion probabilities of the regressors. The latter is especially critical for this literature, as the relative importance of the regressors as drivers of growth is often the key motivation for the analysis.

By repeatedly splitting the samples into an inference part and a prediction part, we also examine the robustness of the inference with respect to changes to the data set and we assess the predictive performance of the model with the various prior settings.

The next section briefly summarizes the main ideas of BMA. Section 3 describes the Bayesian model, and Section 4 examines the theoretical consequences of the priors in more detail. Empirical results for three data sets are provided in Section 5 (using the full samples) and Section 6 (using 100 randomly generated subsamples of a given size). The final section concludes and provides recommendations for users of BMA in this literature.

2. Bayesian Model Averaging in a Nutshell

There exist many excellent and detailed descriptions of BMA in the literature—*e.g.*, Hoeting *et al.* (1999), or Chipman *et al.* (2001)—but we will briefly survey the main issues here for the sake of completeness.

In the face of model uncertainty, a formal Bayesian approach is to treat the model index as a random variable, and to use the data to conduct inference on it. Let us assume that in order to describe the data y we consider the possible models $M_j, j = 1, \dots, J$, grouped in the model space $\mathcal{M}$. In order to give a full probabilistic description—*i.e.*, a Bayesian model—of the problem, we now need to specify a prior $P(M_j)$ on $\mathcal{M}$ and the data will then lead to a posterior $P(M_j \mid y)$.

This posterior can be used to simply select the “best” model (usually the one with highest posterior probability). However, in the case where posterior mass on $\mathcal{M}$ is not strongly concentrated on a particular model, it would not be wise to leave out all others. The strategy of using only the best model has been shown to predict worse than BMA, which mixes over models, using the posterior model probabilities as weights. The superior predictive performance of these methods has been established using both decision theory and empirical results—see, *e.g.*, Min and Zellner, (1993), Raftery *et al.* (1997).

Thus, under BMA inference on some quantity of interest, Δ, which is not model-specific, such as a predictive quantity or the effect of some covariate, will then be obtained through mixing the inference from each individual model

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

This implies a fully probabilistic treatment of model uncertainty, just like parameter uncertainty.

In order to implement BMA in practice, we thus need to be able to compute the posterior distribution on $\mathcal{M}$. It follows directly from Bayes' Theorem that $P(M_j \mid y) \propto l_y(M_j)P(M_j)$, where $l_y(M_j)$, the marginal likelihood of M_j, is simply the likelihood integrated with the prior on the parameters of M_j, denoted here by $p(\theta_j \mid M_j)$. Thus,

$$l_y(M_j) = \int p(y \mid \theta_j, M_j)\, p(\theta_j \mid M_j)\, d\theta_j. \tag{2}$$

A further complication can occur when J is large, making exhaustive computation of the sum in (1) prohibitively expensive in computational effort. Then we often resort to simulation over $\mathcal{M}$. In particular, we use a Markov chain Monte Carlo (MCMC) sampler to deal with the very large model space $\mathcal{M}$ (already containing 2.2×10^{12} models for the smallest example here with $k = 41$). Since the posterior odds between any two models are analytically available (see Subsection 4.3), this sampler moves in model space alone. Thus, the MCMC algorithm is merely a tool to deal with the practical impossibility of exhaustive analysis of $\mathcal{M}$, by only visiting the models which have non-negligible posterior probability. In this paper we use the Fortran code from FLS updated to account for data sets with more than 52 regressors, as explained in Ley and Steel (2007). This code can deal with up to 104 regressors, corresponding to a model space $\mathcal{M}$ containing $2^{104} = 2 \times 10^{31}$ models, and is available at the *Journal of Applied Econometrics* data and code archive.

Other approaches to dealing with the large model space are the use of a coin-flip importance-sampling algorithm in SDM, and the branch-and-bound method developed by Raftery (1995). Eicher *et al.* (2007) experiment with all three algorithms on the FLS data and find that results from the MCMC and branch-and-bound methods are comparable, with the coin-flip method taking substantially more computation time, and leading to somewhat different results.

3. The Bayesian Model

In keeping with the literature, we adopt a Normal linear regression model for n observations of growth in per capita GDP, 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 variables in Z to appear in the model. This results in 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

$$y \mid \alpha, \beta_j, \sigma \sim \mathrm{N}(\alpha \iota_n + Z_j \beta_j, \sigma^2 I), \tag{3}$$

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

$$p(\alpha, \beta_j, \sigma \mid M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j|0, \sigma^2(gZ_j'Z_j)^{-1}), \tag{4}$$

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. Of course, we need a proper prior on β_j in (4), as an improper prior would not allow for meaningful Bayes factors. The general prior structure in (4), sometimes with small changes, is shared by many papers in the growth regression literature, and also in the more general literature on covariate selection in linear models—see, *e.g.*, Clyde and George (2004) for a recent survey.

Based on theoretical considerations and extensive simulation results in Fernández *et al.* (2001a), FLS choose to use $g = 1/\max\{n, k^2\}$ in (4). In the sequel, we shall mainly focus on the two choices for g that underlie this recommendation.

[1] The first choice, $g_{0j} = 1/n$, roughly corresponds to assigning the same amount of information to the conditional prior of β as is contained in one observation. Thus, it is in the spirit of the "unit information priors" of Kass and Wasserman (1995) and the original g-prior used in Zellner and Siow (1980). Fernández *et al.* (2001a) show that log Bayes factors using this prior behave asymptotically like the Schwarz criterion (BIC), and George and Foster (2000) show that for known σ^2 model selection with this prior exactly corresponds to the use of BIC.

[2] The second choice is $g_{0j} = 1/k^2$, which is suggested by the Risk Inflation Criterion of Foster and George (1994). In growth regression, we typically have that $k^2 \gg n$ (as is the case in all three examples here), so that the recommendation of Fernández *et al.* (2001a) would lead to the use of $g = 1/k^2$.

In this paper, we shall not consider other choices for g, but some authors suggest making g random—*i.e.*, putting a hyperprior on g. In fact, the original Zellner-Siow prior can be interpreted as such, and Liang *et al.* (2005) propose the class of hyper-g priors, which still allow for closed form expressions for the marginal likelihood in (2).

The prior model probabilities are often specified by $P(M_j) = \theta^{k_j}(1-\theta)^{k-k_j}$, assuming that each regressor enters a model independently of the others with prior probability θ. Raftery *et al.* (1997), Fernández *et al.* (2001a) and FLS choose $\theta = 0.5$, which can be considered a benchmark choice—implying that $P(M_j) = 2^{-k}$ and that expected model size is $k/2$. SDM examine the sensitivity of their results to the choice of θ. The next section will consider the prior on the model space $\mathcal{M}$ more carefully.

The assumption of prior independent inclusion of regressors can be contentious in some contexts. Chipman *et al.* (2001) argue that in situations where interactions are considered, or some covariates are collinear, it may be counterintuitive to treat the inclusion of each regressor as independent a priori. In particular, they recommend "dilution" priors, where model probabilities are diluted across neighbourhoods of *similar* models. From an economic perspective, a related idea was proposed by Brock *et al.* (2003), who construct the model prior by focusing on economic theories rather than individual regressors. This implies a hierarchical tree structure for the prior on model space and was also used, *e.g.*, in Durlauf *et al.* (2006). In this paper, we will only consider priors that assign equal probabilities to the inclusion of each variable, as this is still by far the most common practice and requires less elicitation effort on the part of the user. In that sense, it is closer to the idea of a 'non-informative' prior, which is our main focus here.

4. Prior Assumptions and Posterior Inference

4.1. Model prior specification and model size

In order to specify a prior on model space, consider the indicator variable γ_i, which takes the value 1 if covariate i is included in the regression and 0 otherwise, $i = 1, \ldots, k$. Given the probability of inclusion, say θ, γ_i will then have a Bernoulli distribution: $\gamma_i \sim \text{Bern}(\theta)$, and if the inclusion of each covariate is independent then the *model size* W will have a Binomial distribution:

$$W \equiv \sum_{i=1}^{k} \gamma_i \sim \text{Bin}(k, \theta).$$

This implies that, if we fix θ—as was done, *e.g.*, in FLS and SDM, as in most other studies—the prior model size will have mean θk and variance $\theta(1-\theta)k$.

Typically, the use of a hierarchical prior increases the flexibility of the prior and reduces the dependence of posterior and predictive results (including model probabilities) on prior assumptions. Thus, making θ random rather than fixing it would seem a sensible extension—this was implemented by Brown *et al.* (1998), and is also discussed in, *e.g.*, Clyde and George (2004) and Nott and Kohn (2005). An obvious choice for the distribution of θ is a Beta with hyperparameters $a, b > 0$, *i.e.*, $\theta \sim \text{Be}(a, b)$, leading to the following prior moments for model size, as a function of k, a and b:

$$\text{E}[W] = \frac{a}{a+b} k, \tag{5}$$

$$\text{Var}[W] = \frac{ab(a+b+k)}{(a+b)^2(a+b+1)} k. \tag{6}$$

This framework generates a prior model size distribution that corresponds to the Binomial-Beta distribution (Bernardo and Smith, 1994, p. 117), and has a probability mass function given by

$$P(W = w) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+k)} \binom{k}{w} \Gamma(a+w)\,\Gamma(b+k-w), \quad w = 0, \ldots, k.$$

In the special case where $a = b = 1$ (*i.e.*, we mix with a uniform prior for θ) we obtain a discrete uniform prior for model size with $P(W = w) = 1/(k+1)$ for $w = 0, \ldots, k$.

This prior depends on two parameters, (a, b), and it will facilitate prior elicitation to fix $a = 1$. This allows for a wide range of prior behaviour and generally leads to reasonable prior assumptions, as seen below. It is attractive to elicit the prior in terms of the prior mean model size, m. The choice of $m \in (0, k)$ will then determine b through equation (5), which implies $b = (k-m)/m$.

Thus, in this setting, the analyst only needs to specify a prior mean model size, which is exactly the same information one needs to specify for the case with fixed θ, which should then equal

$\theta = m/k$. With this Binomial-Beta prior, the prior mode for W will be at zero for $m < k/2$ and will be at k for $m > k/2$. The former situation is likely to be of most practical relevance, and, in that case, the prior puts most mass on the null model, which reflects a mildly conservative prior stance, where we require some data evidence to favour the inclusion of regressors.

For the case of $k = 67$, Figure 1 contrasts the prior model-size distributions[1] with fixed θ (solid lines) and random θ (dashed lines), for two choices for mean model size: $m = 7$, which is used in SDM, and $m = 33.5$, which corresponds to a uniform prior in the random θ case. Clearly, the prior with fixed θ is very far from uniform, even for $m = k/2$. Generally, the difference between the fixed and random θ cases is striking: prior model size distributions for fixed θ are very concentrated.

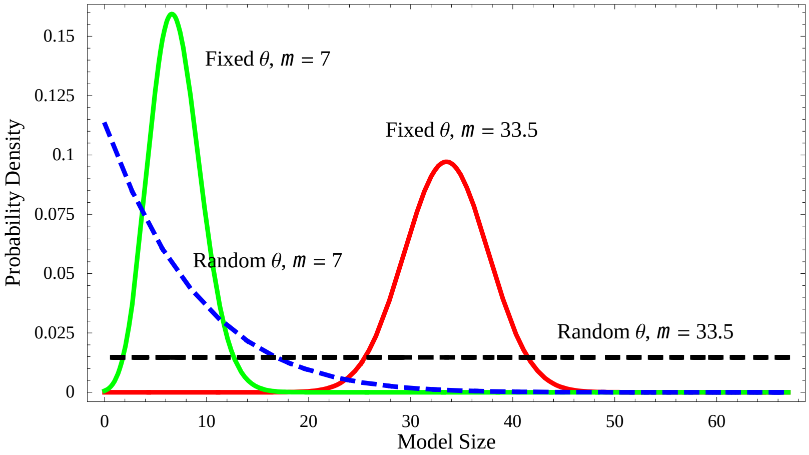

Fig. 1. Prior model size for $k = 67$; fixed θ (solid) and random θ (dashed).

With random θ and the hyperparameter choices discussed above, the prior variance of model size is then equal to

$$\mathrm{Var}[W] = m(m+1)\frac{k-m}{k+m}, \tag{7}$$

which is roughly equal to m^2 when $k \gg m$. In fact, we can illustrate the added uncertainty introduced by the hierarchical prior structure on θ by the fact that the variance in (7) multiplies $\mathrm{Var}[W]$ of the corresponding case with fixed $\theta = m/k$ by a factor

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

which, for $k > 1$ is always larger than one and is an increasing function of m, ranging from 1 (in the limit as $m \downarrow 0$) to $(k+1)/2$ (in the limit as $m \uparrow k$).

Thus, treating θ as random will typically imply a substantial increase in the prior uncertainty about model size. To assess whether this increase is reasonable, consider Pearson's coefficient of

[1] These distributions are, of course, discrete, but for ease of presentation they are depicted through continuous graphs—the same comment applies throughout the paper.

variation, say CV$[W]$, given by the standard deviation divided by the mean. For fixed $\theta = k/m$ this coefficient of variation equals

$$\mathrm{CV}[W] = \sqrt{\frac{k-m}{km}},$$

which is a rapidly decreasing function of m and is unity for $m = k/(k+1)$, which is often close to one. Thus, for any reasonable prior mean model size, the prior with fixed θ will be far too tight. For example, if we take $m = 7$ in our applications, where k ranges from 41 to 67, CV$[W]$ will range from 0.344 to 0.358, which is quite small. For $m = k/2$, then CV$[W] = \sqrt{1/k}$, which ranges from 0.122 to 0.156, clearly reflecting an unreasonable amount of precision in the prior model size distribution.

For random θ with the hyperprior as described above, we obtain

$$\mathrm{CV}[W] = \sqrt{\frac{(m+1)(k-m)}{m(k+m)}},$$

which is also decreasing in m, but is much flatter than the previous function over the range of practically relevant values for m.[2] Taking $m = 7$ in our applications, CV$[W]$ will now range from 0.900 to 0.963, which is much more reasonable. For $m = k/2$, we now have that CV$[W] = \sqrt{(k+2)/3k}$, which ranges from 0.586 to 0.591.

Thus, this hierarchical prior seems quite a sensible choice. In addition, both Var$[W]$ and CV$[W]$ increase with k for a given m, which also seems a desirable property. This holds for both prior settings; however, $\lim_{k\to\infty}$ CV$[W] = \sqrt{(m+1)/m}$ for the case with random θ whereas this limit is only $\sqrt{1/m}$ for the fixed θ case.

4.2. Prior odds

Posterior odds between any two models in $\mathcal{M}$ are given by

$$\frac{P(M_i|y)}{P(M_j|y)} = \frac{P(M_i)}{P(M_j)} \cdot \frac{l_y(M_i)}{l_y(M_j)},$$

where $l_y(M_i)$ is the marginal likelihood, defined in (2). Thus, the prior distribution on model space only affects posterior model inference through the prior odds ratio $P(M_i)/P(M_j)$. For a prior with a fixed $\theta = 0.5$ prior odds are equal to one (*i.e.*, each model is a priori equally probable). If we fix θ at a different value, these prior odds are

$$\frac{P(M_i)}{P(M_j)} = \left(\frac{\theta}{1-\theta}\right)^{k_i - k_j},$$

[2] Of course, both CV functions tend to zero for $m \uparrow k$ since then all prior mass has to be on the full model. In addition, CV in both cases tends to ∞ as $m \downarrow 0$.

thus inducing a prior penalty for the larger model if $\theta < 0.5$ and favouring the larger model for values of $\theta > 0.5$. Viewed in terms of the corresponding mean model size, m, we obtain (for $\theta = m/k$):

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

from which it is clear that the prior favours larger models if $m > k/2$. For the hierarchical Be(a, b) prior on θ, we obtain the prior model probabilities:

$$P(M_j) = \int_0^1 P(M_j|\theta)p(\theta)\mathrm{d}\theta = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+k_j)\Gamma(b+k-k_j)}{\Gamma(a+b+k)}.$$

Using $a = 1$ and our prior elicitation in terms of E$[W] = m$ as above, we obtain the following prior odds

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

Figure 2 compares the log prior odds induced by the fixed and random θ prior structures, in the situation where $k = 67$ and for $m = 7, 33.5$ and 50. The graphs indicate the prior odds in favour of a model with $k_i = 10$ versus models with varying k_j.

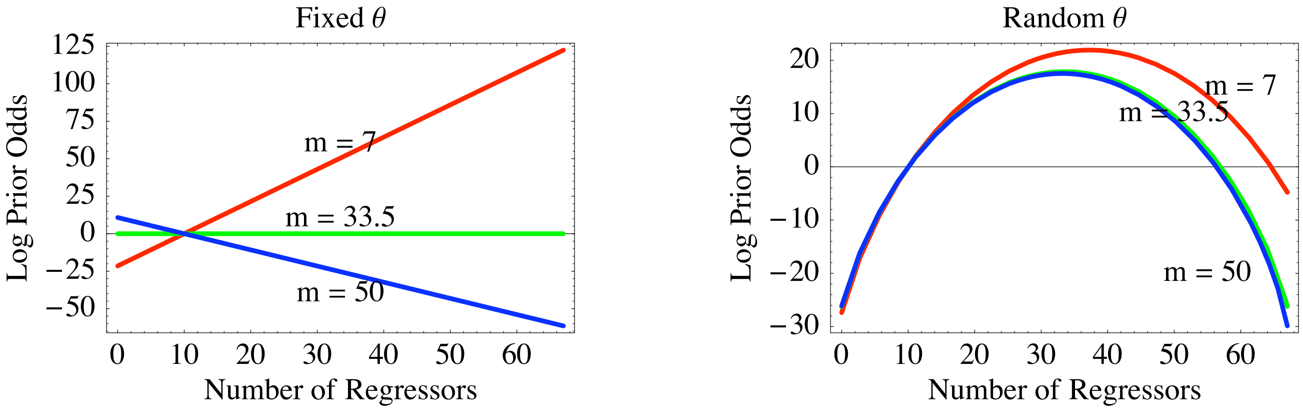

Fig. 2. Log of Prior Odds: $k_i = 10$ vs varying k_j.

Note that the random θ case always leads to downweighting of models with k_j around $k/2$, irrespectively of m. This counteracts the fact that there are many more models with k_j around $k/2$ in the model space $\mathcal{M}$ than for k_j values nearer to 0 or k.[3] In contrast, the prior with fixed θ does not take the number of models at each k_j into account and simply always favours larger models when $m > k/2$ and the reverse when $m < k/2$. Note also the wider range of values that the log prior odds take in the case of fixed θ.

[3] The number of models with k_j regressors in $\mathcal{M}$ is given by $\binom{k}{k_j}$. For example, with $k = 67$, we have 1 model with $k_j = 0$ and $k_j = k$, 8.7×10^8 models with $k_j = 7$ and $k_j = 60$ and a massive 1.4×10^{19} models with $k_j = 33$ and 34.

Thus, the choice of m is critical for fixed θ, but much less so for random θ. The latter prior structure is naturally adaptive to the data observed. This is, again, illustrated by Figure 3, which plots the log of the prior odds of a model with $k_i = (k_j - 1)$ regressors versus a model with k_j regressors as a function of k_j.

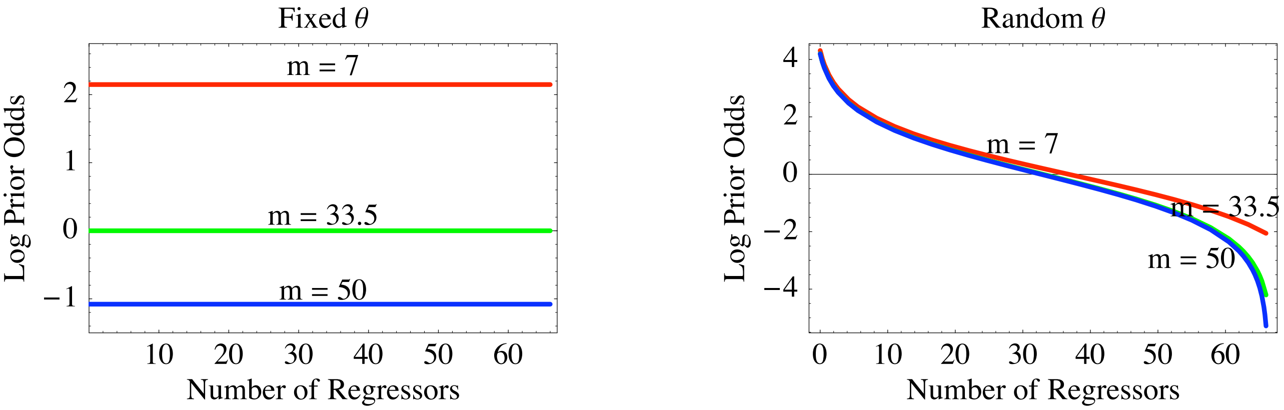

Fig. 3. Log of Prior Odds: $k_i = (k_j - 1)$ vs varying k_j.

Whereas the fixed θ prior always favours the smaller model M_i for $m < k/2$, the choice of m for random θ only moderately affects the prior odds, which swing towards the larger model when k_j gets larger than approximately $k/2$. This means that using the prior with fixed θ will have a deceptively strong impact on posterior model size. This prior does not allow for the data to adjust prior assumptions on mean model size that are at odds with the data, making it a much more risky choice.

4.3. Bayes factors

The marginal likelihood in (2) forms the basis for the Bayes factor (ratio of marginal likelihoods) and can be derived analytically for each model with prior structure (4) on the model parameters. Provided g in (4) does not depend on the model size k_j, the Bayes factor for any two models from (3)–(4) becomes:

$$\frac{l_y(M_i)}{l_y(M_j)} = \left(\frac{g}{g+1}\right)^{\frac{k_i-k_j}{2}} \left(\frac{1+g-R_i^2}{1+g-R_j^2}\right)^{-\frac{n-1}{2}}, \tag{8}$$

where R_i^2 is the usual coefficient of determination for model M_i, *i.e.*, $R_i^2 = 1 - [y'Q_{X_i}y/(y - \bar{y}\iota_n)'(y - \bar{y}\iota_n)]$ and we have defined $Q_A = I - A(A'A)^{-1}A'$ and $X_i = (\iota_n, Z_i)$, the design matrix of M_i, which is always assumed to be of full column rank. The expression in (8) is the relative weight that the data assign to the corresponding models, and depends on sample size n, the factor g of the g-prior and the size and fit of both models, with the latter expressed through R^2.

Let us compare this with the so-called BACE approach of SDM. The BACE approach is not totally Bayesian, as it is not formally derived from a prior-likelihood specification, but relies on an approximation as sample size, n, goes to infinity (which may not be that realistic in the growth context). In fact, BACE uses the Schwarz approximation to compute the Bayes factor, as was earlier used in Raftery (1995) in a very similar context. From equation (6) in SDM, we get the following Bayes factor:

$$\frac{l_y(M_i)}{l_y(M_j)} = n^{\frac{k_j - k_i}{2}} \left(\frac{1 - R_i^2}{1 - R_j^2} \right)^{-\frac{n}{2}} . \tag{9}$$

This expression is not that different from the one in our equation (8), provided we take $g = 1/n$. In that case, the Bayes factor in (8) becomes:

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

which behaves very similarly to the BACE procedure in (9) for practically relevant values of n (as in the examples here). This will be crucial in explaining the similarity of the results with the FLS and SDM prior settings mentioned before.

It also becomes immediately clear that the necessity of choosing prior settings implicit in using BMA is not really circumvented by the use of BACE, in contrast with the claims in SDM. In fact, BACE implicitly fixes g in the context of our BMA framework. The fact that this is hidden to the analyst does not make the results more robust with respect to this choice. It even carries a substantial risk of conveying a false sense of robustness to the applied analyst.

Now we can examine more in detail how the various prior choices translate into model size penalties. From (8) we immediately see that if we have two models that fit equally well (*i.e.*, $R_i^2 = R_j^2$), then the Bayes factor will approximately equal $g^{(k_i - k_j)/2}$ (as g tends to be quite small). If one of the models contains one more regressor, this means that the larger model will be penalized by $g^{1/2}$.

For $n = 88$ and $k = 67$ (as in the SDM data) this means that the choice of $g = 1/n$ leads to a Bayes factor of 0.107 and choosing $g = 1/k^2$ implies a Bayes factor of 0.015. Thus, the model size penalty is much more severe for $g = 1/k^2$ in the context of these types of data. The size penalty implicit in the BACE procedure is the same as for $g = 1/n$.

We can also ask how much data evidence is required to exactly compensate for the effects of prior odds under different specifications. Posterior odds will be unity if the Bayes factor equals the inverse of the prior odds, thus if

$$\frac{l_y(M_i)}{l_y(M_j)} = \frac{P(M_j)}{P(M_i)}, \tag{10}$$

where the prior odds are a function of prior mean model size m, as well as k and the sizes of both models—as described in the previous subsection.

Typically, we have no control over n and k, but we do need to select g and m. To give more insight into the tradeoff between g and m, Figure 4 plots the contours in (g, m)-space that correspond to $n = 88, k = 67, k_i = 8, k_j = 7$, and $R_j^2 = 0.75$ for different ratios R_i^2/R_j^2 that would make the two models equally probable in the posterior. These plots are provided for fixed and random θ model priors with the Bayes factor in (6).

It is apparent from the left figure for the fixed θ case that there is a clear trade-off between m and g: larger m, inducing a smaller size penalty, can be compensated by a small g, which increases

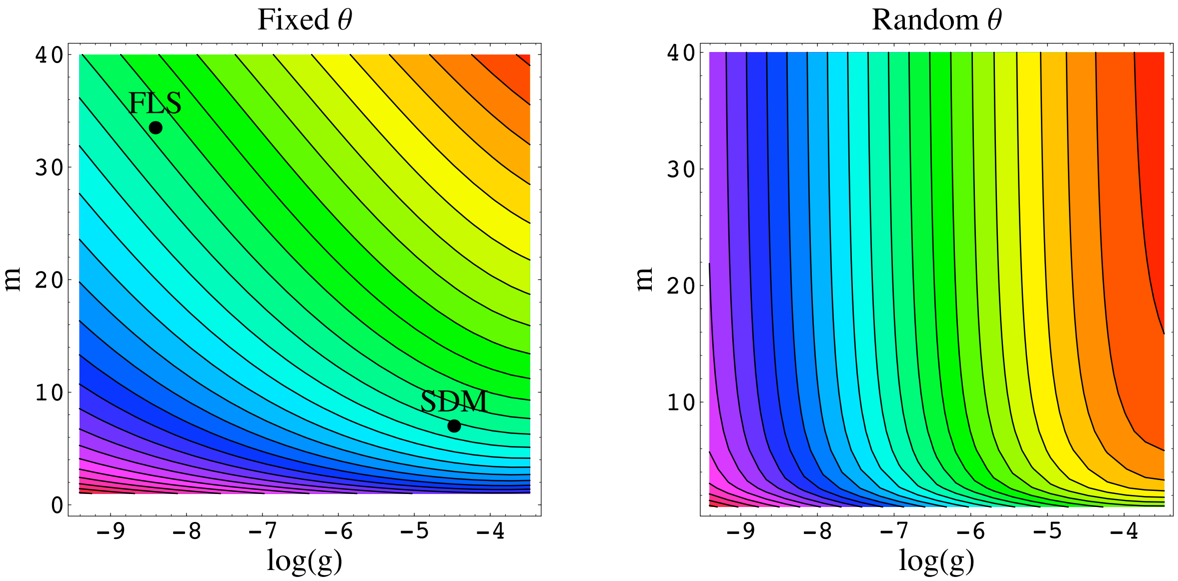

Fig. 4. Equal Probability Contours for different ratios R_i^2/R_j^2.
Left Panel shows also the choices of $(\log(g), m)$ for FLS and SDM.
($n = 88, k = 67, k_i = 8, k_j = 7$, and $R_j^2 = 0.75$)

the penalty. In this particular case with k and n corresponding to the SDM data, we notice that the combination of $g = 1/n$ with $m = 7$, as implicitly assumed in SDM, is on a contour rather close to the one corresponding to the settings in FLS: $g = 1/k^2$ with $m = 33.5$. These combinations are represented in the left panel. This explains the similarity of the results with the SDM and FLS prior settings, mentioned in the Introduction. See also the related discussion in Subsection 3.2 of Eicher *et al.* (2007). If we adopt the hierarchical prior with random θ, the trade-off between g and m almost disappears, as the choice of m now only has a very small role to play.

5. Some Illustrative Examples: Complete Sample Results

In this section we will present posterior results for three data sets that have been analysed in the literature—all the datasets and the code used here are available on the *Journal of Applied Econometrics* data and code archive. The results reported in this section are based on MCMC runs of 2 million retained drawings after a burn-in of 1 million.

We focus, in particular, on the effect of the prior choices on posterior model size distributions, the spread of the posterior mass over model space, posterior model probabilities and the inclusion of individual regressors.

Results are presented for eight combinations of prior settings, taking $g = 1/n$ and $g = 1/k^2$, $m = 7$ and $m = k/2$ and using either a fixed or a random θ. The choice of $m = 7$ was used in SDM and also motivated by the restriction to models with up to 7 regressors in Levine and Renelt (1992) and exactly 7 regressors in Sala-i-Martin (1997). Choosing $m = k/2$ corresponds to a "vague" model size prior, which is, as discussed in Subsection 4.1, uniform for random θ, and symmetric (but far from uniform) for fixed θ.

5.1. *The FLS Data*

We first illustrate the effects of our prior choices using the growth data of FLS. The latter data set contains k = 41 potential regressors to model the average per capita GDP growth over 1960-1992 for a sample of n = 72 countries.

As expected, results with $m = \frac{41}{2}$ = 20.5, fixed θ and $g = 1/k^2$ (the FLS settings) are virtually identical to those obtained in FLS on the basis of a chain of the same length. Figure 5 provides a complete picture of the posterior model size distributions (overplotted with the priors) for all eight combinations of prior settings. Summaries in the form of the first two moments are provided in Table 1.

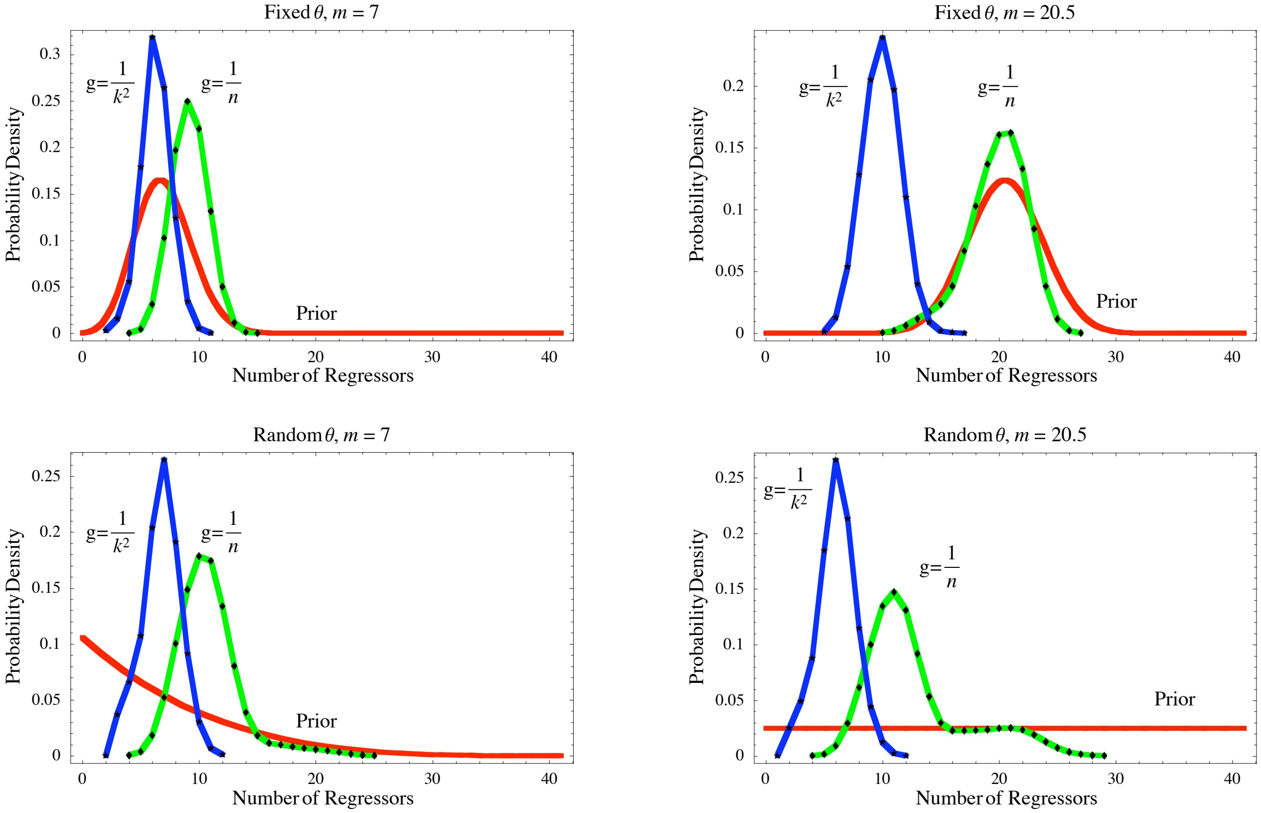

Fig. 5. Model Size: Prior and Posterior Distributions for FLS data.

From these results, we immediately notice the striking influence that prior assumptions have on model size in the posterior. Even if we fix the prior mean model size, simply changing between random and fixed θ, or choosing a different g can substantially alter the posterior distribution of model size, W. Clearly, influence of the choice of m is massive for fixed θ, whereas its effect is much less severe for the case of random θ. This accords with the fact (discussed in Subsection 4.1) that the prior on model size is very concentrated for the fixed θ case, and the relative robustness with respect to the choice of m noted for random θ in Subsections 4.2 and 4.3. Also, it appears that the effects of prior choices are much more pronounced for the prior with $g = 1/n$. In fact, if we adopt $g = 1/k^2$ with a hierarchical prior on θ, the value chosen for m has very little impact, in

Table 1. FLS Data—prior and posterior moments of model size.
Properties of the chain and the best model.

	m = 7				m = $k/2$ = 20.5			
	fixed θ		random θ		fixed θ		random θ	
	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$
Prior mean	7	7	7	7	20.5	20.5	20.5	20.5
Prior st. dev.	2.41	2.41	6.30	6.30	3.20	3.20	12.12	12.12
Posterior mean	9.17	6.29	10.76	5.73	19.84	9.91	12.87	6.03
Posterior st. dev.	1.54	1.30	2.67	1.65	2.57	1.63	4.43	1.64
# Models visited	141,980	20,003	329,694	19,275	522,637	150,255	436,602	23,558
# Models covering 50%	1,461	40	3,569	27	9,348	1,598	5,489	36
Post. Prob. best model	0.91%	6.22%	0.65%	5.61%	0.21%	1.24%	0.54%	5.21%
k_j for best model	9	6	10	5	20	10	10	6
Corr. visits and PO	0.992	0.999	0.971	0.999	0.748	0.992	0.922	0.998
Prob. mass visited	70.7%	95.2%	39.6%	95.1%	18.4%	69.7%	27.2%	93.7%

marked contrast with the other prior settings.

From Figure 5 we also deduce that the posterior of W can display bimodality if data and prior information are in conflict. This happens when m = 20.5 and we choose $g = 1/n$ for the case with random θ. In the corresponding fixed θ case the prior is actually so concentrated that it dominates and the posterior is mostly determined by the prior.

Table 1 also records some key properties of the chain: it is clear that the choice of $g = 1/n$ always leads to the sampler visiting many more models, indicating that the posterior mass is more spread out over the model space. This is clear when considering how many of the higher-probability models we need to cover 50% of the probability mass. This number varies dramatically with the choice of g, especially for random θ (where we need more than 100 times as many models for $g = 1/n$). If we adopt $g = 1/n$, many larger models are visited and the posterior mass assigned to the best model is much smaller than with $g = 1/k^2$ (roughly by a factor 8). In addition, the best model is much larger for $g = 1/n$. All this is in keeping with the smaller penalty for increasing model size that is implied by adopting $g = 1/n$ (see Subsection 4.3). An important practical consequence of this is that convergence of the chain requires a longer run. Table 1 also presents the correlation between model visit frequencies and probabilities computed on the basis of the exact posterior odds of the visited models, which is in excess of 0.99 (indicating excellent convergence) for all cases with $g = 1/k^2$. For $g = 1/n$, where the model probability is spread more evenly over $\mathcal{M}$, the evidence in favour of convergence is slightly less convincing. For the case with m = 20.5 and fixed θ we would ideally recommend a longer run.[4] However, even in this case, multiple runs led to very similar findings. For completeness, Table 1 also displays the estimated total posterior model probability visited by the chain, computed as suggested in George and McCulloch (1997).

For the random θ cases, the specification and the posterior probability of the best model is not

[4] It is important for our purposes in this section that run lengths are identical, to ensure comparability of the properties of the chains. Longer chains will typically capture marginally more of the posterior probability, but they will only add models with very small posterior probabilities and this will not affect any of the conclusions.

much affected by the choice of m. However, changing from $g = 1/n$ to $g = 1/k^2$ has a dramatic effect on both. Finally, the size of the best model varies in between 5 and 20, and is not much affected by m for random θ.

Of course, one of the main reasons for using BMA in the first place is to assess which of the regressors are important for modelling growth. Table 2 presents the marginal posterior inclusion probabilities of all regressors that receive an inclusion probability of over 10% under any of the prior settings. It is clear that there is a large amount of variation in which regressors are identified as important, depending on the prior assumptions.

Whereas three variables (past GDP, fraction Confucian and Equipment investment) receive more than 0.75 inclusion probability and a further two (Life expectancy and the Sub-Saharan dummy) are included with at least probability 0.50 in all cases, there are many differences. If we compare cases that only differ in m, the choice of $g = 1/n$ with fixed θ leads to dramatic differences in inclusion probabilities: Fraction Hindu, the Labour force size, and Higher education enrollment go from virtually always included with m = 20.5 to virtually never included with m = 7; the Number of years open economy has the seventh largest inclusion probability for m = 7 and drops to the bottom of the 32 variables shown in the table for m = 20.5. In sharp contrast, the case with $g = 1/k^2$ and random θ leads to very similar inclusion probabilities for both values of m.

Finally, note that results for model size, chain behaviour and inclusion probabilities are quite similar for the cases where $g = 1/n$ with fixed $\theta = \frac{7}{41}$ (the preferred implied prior in SDM) and where $g = 1/k^2$ with $\theta = 0.5$ (the prior used in FLS). This is in line with the negative trade-off between g and m illustrated in Figure 4, which explains the similarity between empirical results using BACE and the FLS prior on the same data. The same behaviour is observed for the other datasets presented in the next subsections.

5.2. The Data Set of MP

MP investigate the role of initial conditions at independence from colonial rule on the economic growth of African countries. These authors focus on the average growth rate in GDP from 1960 to 1992 and construct a dataset for n = 93 countries with k = 54 covariates, obtained by combining 32 original regressors and 22 interaction dummies.

Table 3 records the main characteristics of model size and the MCMC chain, and illustrates that results are quite close to those with the FLS data, leading to the same conclusions. Probably due to the somewhat larger model space, convergence is now even more problematic for the fixed θ case with $g = 1/n$ and $m = \frac{54}{2} = 27$. Again, however, longer runs do not lead to appreciably different conclusions (see footnote 5).

5.3. The Data Set of SDM and DW

SDM and Doppelhofer and Weeks (2008) use a larger data set, and model annual GDP growth per capita between 1960 and 1996 for n = 88 countries as a function of k = 67 potential drivers.

Despite the larger model space (the number of models in $\mathcal{M}$ is now 1.5×10^{20}), Table 4 shows that posterior model probabilities are more concentrated than in the previous two cases; in fact, even in those cases where many models are visited, 50% of the posterior mass is still accounted for through a rather small number of models. Also, model sizes tend to be smaller; in fact, the posterior

Table 2. FLS data—Marginal posterior inclusion probabilities of the covariates.

	m = 7				$m = k/2$ = 20.5			
	fixed θ		random θ		fixed θ		random θ	
Regressors	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$
log GDP in 1960	1.00	0.91	1.00	0.79	1.00	1.00	1.00	0.84
Fraction Confucian	0.99	0.94	1.00	0.93	1.00	1.00	1.00	0.94
Life expectancy	0.92	0.74	0.95	0.63	1.00	0.95	0.97	0.68
Equipment investment	0.95	0.98	0.93	0.98	0.97	0.94	0.94	0.98
Sub-Saharan dummy	0.70	0.59	0.78	0.53	1.00	0.75	0.85	0.55
Fraction Muslim	0.62	0.29	0.63	0.23	0.43	0.66	0.61	0.27
Rule of law	0.41	0.17	0.56	0.15	0.93	0.52	0.67	0.17
Number of years open economy	0.56	0.60	0.46	0.54	0.07	0.50	0.36	0.56
Degree of capitalism	0.36	0.09	0.50	0.08	0.56	0.47	0.56	0.09
Fraction Protestant	0.38	0.23	0.49	0.24	0.47	0.46	0.51	0.24
Fraction GDP in mining	0.36	0.08	0.51	0.07	0.94	0.44	0.63	0.08
Non-Equipment investment	0.33	0.07	0.47	0.06	0.71	0.43	0.56	0.07
Latin American dummy	0.18	0.09	0.23	0.07	0.75	0.19	0.34	0.08
Primary school enrollment, 1960	0.19	0.10	0.21	0.08	0.63	0.18	0.29	0.09
Fraction Buddhist	0.14	0.05	0.20	0.07	0.25	0.17	0.23	0.07
Black market premium	0.11	0.02	0.22	0.01	0.69	0.16	0.34	0.02
Fraction Catholic	0.09	0.03	0.13	0.02	0.12	0.11	0.14	0.03
Civil liberties	0.09	0.03	0.13	0.02	0.54	0.10	0.22	0.03
Fraction Hindu	0.06	0.07	0.18	0.01	0.97	0.10	0.36	0.01
Political rights	0.06	0.01	0.09	0.01	0.29	0.07	0.13	0.02
Exchange rate distortions	0.06	0.03	0.06	0.02	0.12	0.06	0.08	0.03
Age	0.06	0.02	0.07	0.02	0.25	0.06	0.10	0.02
War dummy	0.05	0.02	0.06	0.02	0.14	0.05	0.08	0.02
Fraction of Pop. Speaking English	0.04	0.01	0.07	0.01	0.42	0.05	0.15	0.01
Size labor force	0.04	0.01	0.11	0.01	0.95	0.05	0.28	0.01
Ethnolinguistic fractionalization	0.03	0.01	0.09	0.01	0.87	0.03	0.24	0.01
Spanish Colony dummy	0.03	0.01	0.06	0.01	0.59	0.03	0.17	0.01
French Colony dummy	0.03	0.01	0.05	0.01	0.54	0.03	0.15	0.01
Higher education enrollment	0.02	0.01	0.08	0.01	0.91	0.02	0.24	0.01
British colony dummy	0.02	0.00	0.04	0.00	0.47	0.02	0.13	0.00
Outward orientation	0.02	0.01	0.04	0.01	0.42	0.02	0.12	0.01
Public education share	0.02	0.00	0.03	0.00	0.30	0.02	0.08	0.00

mean model size is less than 3 for three of the four cases with $g = 1/k^2$. If we adopt a random θ and $g = 1/k^2$, the choice for m has little effect. The twenty best models are the same for both values of m, with very similar posterior probabilities and only slight differences in ordering. The best model (with over 60% posterior probability) in these cases as well as with two other settings, is the model with only the East Asian dummy and Malaria prevalence as regressors. The same model is also second best for the random θ case with m = 33.5 and $g = 1/n$, but is well down the ordering (receiving less than 0.25% of probability) if we fix θ with m = 33.5. In fact, in the latter case with $g = 1/n$, the regressors East Asian dummy and Malaria prevalence never appear together in any model with posterior probability over 0.25%. This further illustrates the huge impact of simply changing between the fixed and random θ cases.

Convergence is excellent, except for the case with fixed θ, m = 33.5 and $g = 1/n$, as in the other examples. The difference in convergence between the cases is even more striking than with the previous examples, and it appears that the case with convergence problems struggles to adequately describe the posterior distribution on the model space $\mathcal{M}$. While most of the mass is covered by a relatively small number of models, the prior assumptions induce a very fat tail of models with

Table 3. MP Data—prior and posterior moments of model size.
Properties of the chain and the best model.

	$m = 7$				$m = k/2 = 27$			
	fixed θ		random θ		fixed θ		random θ	
	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$
Prior mean	7	7	7	7	27	27	27	27
Prior st. dev.	2.47	2.47	6.57	6.57	3.67	3.67	15.88	15.88
Posterior mean	8.68	6.05	9.67	5.42	17.90	9.77	10.37	5.75
Posterior st. dev.	1.52	1.16	2.09	1.53	2.44	1.75	2.24	1.48
# Models visited	106,041	13,103	194,230	11,864	516,479	135,353	241,082	13,88
# Models covering 50%	933	31	1,549	23	3,038	1,353	2,049	28
Post. Prob. best model	1.65%	10.70%	1.14%	9.60%	0.49%	1.28%	0.90%	9.01%
k_j for best model	8	6	8	4	19	8	8	6
Corr. visits and PO	0.990	0.998	0.973	0.999	0.251	0.985	0.963	0.999

Table 4. SDM Data—prior and posterior moments of model size.
Properties of the chain and the best model.

	$m = 7$				$m = k/2 = 33.5$			
	fixed θ		random θ		fixed θ		random θ	
	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$
Prior mean	7	7	7	7	33.5	33.5	33.5	33.5
Prior st. dev.	2.50	2.50	6.74	6.74	4.09	4.09	19.63	19.63
Posterior mean	6.17	2.84	5.26	2.35	14.20	7.00	5.73	2.40
Posterior st. dev.	1.50	0.85	1.86	0.59	1.64	1.44	1.91	0.62
# Models visited	102,852	6,141	101,642	2,167	606,792	130,776	129,929	2,594
# Models covering 50%	590	4	248	1	485	744	383	1
Post. Prob. best model	6.63%	37.72%	5.74%	66.42%	1.43%	6.58%	5.22%	63.64%
k_j for best model	6	2	2	2	13	6	6	2
Corr. visits and PO	0.996	1.000	0.996	1.000	0.029	0.996	0.996	1.000

little but nonnegligible mass. However, inference on most things of interest, such as the regressor inclusion probabilities is not much affected by running longer chains. For all other prior settings, the very large model space is remarkably well explored by the MCMC chain.

6. Robustness and Predictive Analysis: Results from 100 Subsamples

In the previous section we have illustrated that the choice of, perhaps seemingly innocuous, prior settings can have a dramatic impact on the posterior inference resulting from BMA. Posterior model probabilities and identification of the most important regressors can strongly depend on the prior settings we use in our analysis of growth regressions through BMA.

We now address the issue of whether small changes to the data set would result in large changes in inference—*i.e.*, data robustness. In addition, we want to assess the predictive performance of the model under various prior settings. We will use the same device to investigate both issues, namely the partition of the available data into an inference subsample and a prediction subsample. We will then use the various inference subsamples for the evaluation of robustness and assess prediction on the basis of how well the predictive distribution based on the inference subsamples captures the corresponding prediction subsamples.

We take random partitions of the sample, where the size of the prediction subsample is fixed at 15% of the total number of observations (rounded to an integer), leaving 85% of the sample to conduct inference with. We generate random prediction subsamples of a fixed size by using the algorithm of McLeod and Bellhouse (1983). We use 100 random partitions and compute the posterior results through an MCMC chain of 500,000 drawings with a burn-in of 100,000 for each partition. This led to excellent convergence, with the exception of the cases with θ fixed at 0.5 and $g = 1/n$. Thus, for this combination we have used a chain of length 1,000,000 after a burn-in of 500,000, which leads to reliable results.[5] To increase comparability, we use the same partitions for all prior settings.

6.1. Robustness

Figure 6 indicates the distribution of the posterior mean model size across 100 inference subsamples of the FLS data (left panel). Values corresponding to the full sample are indicated by vertical lines. The right panel of the same Figure shows the posterior inclusion probabilities of the ten regressors with highest posterior inclusion probabilities in the full-sample analysis with the FLS prior setting—*i.e.*, $\theta = 0.5$ and $g = 1/\max\{n, k^2\}$.

A striking characteristic of both panels is the sensitivity of the results to the choice of m for the fixed θ priors, whereas the effect of m is very small for the cases with random θ. The choice of g, however, always matters: $g = 1/k^2$ generally leads to smaller mean model sizes and results in very different conclusions on which regressors are important than $g = 1/n$, especially for random θ.

Given each choice of prior, the results vary quite a bit across subsamples, especially for the inclusion probabilities where the interquartile ranges can be as high as 60%. It is interesting how the combinations of $g = 1/n$ with fixed $\theta = 7/41$ (the setting implicitly favoured in SDM) and $g = 1/k^2$ with $\theta = 0.5$ (the prior of FLS) lead to very similar results, both for model sizes and inclusion probabilities, as also noted in Section 5. Overall, the difference between fixed and random θ cases is small for $m = 7$, but substantial for $m = k/2$.

For the MP data, results on posterior model size are quite similar as for the FLS dataset, and we can draw the same conclusions on the basis of the inclusion probabilities.

Finally, Figure 7 shows the robustness results for the SDM data set, where we have $k = 67$ potential regressors. As before, the choice of m critically affects the fixed θ results, but not the ones for random θ. Differences between fixed and random θ cases are large for $m = k/2$, but relatively small for $m = 7$, as with both previous datasets. Again, the choice of g always affects the results, and inference using $g = 1/n$ in combination with $\theta = \frac{7}{67}$ is quite similar to that using $g = 1/k^2$ with $\theta = 0.5$, for both model size and inclusion probabilities. For $g = 1/k^2$ inference on

[5] In fact, the results are very close to those obtained with 100,000 burn-in and 500,000 retained drawings, with the only noticeable difference in the maximum LPS values.

model size is quite concentrated on small values, with the exception of the case with fixed θ and $m = 33.5$.

Even though many results are not very robust with respect to the particular choice of subsample, it is clear that the differences induced by the various prior assumptions largely persist if we take into account small changes to the data.

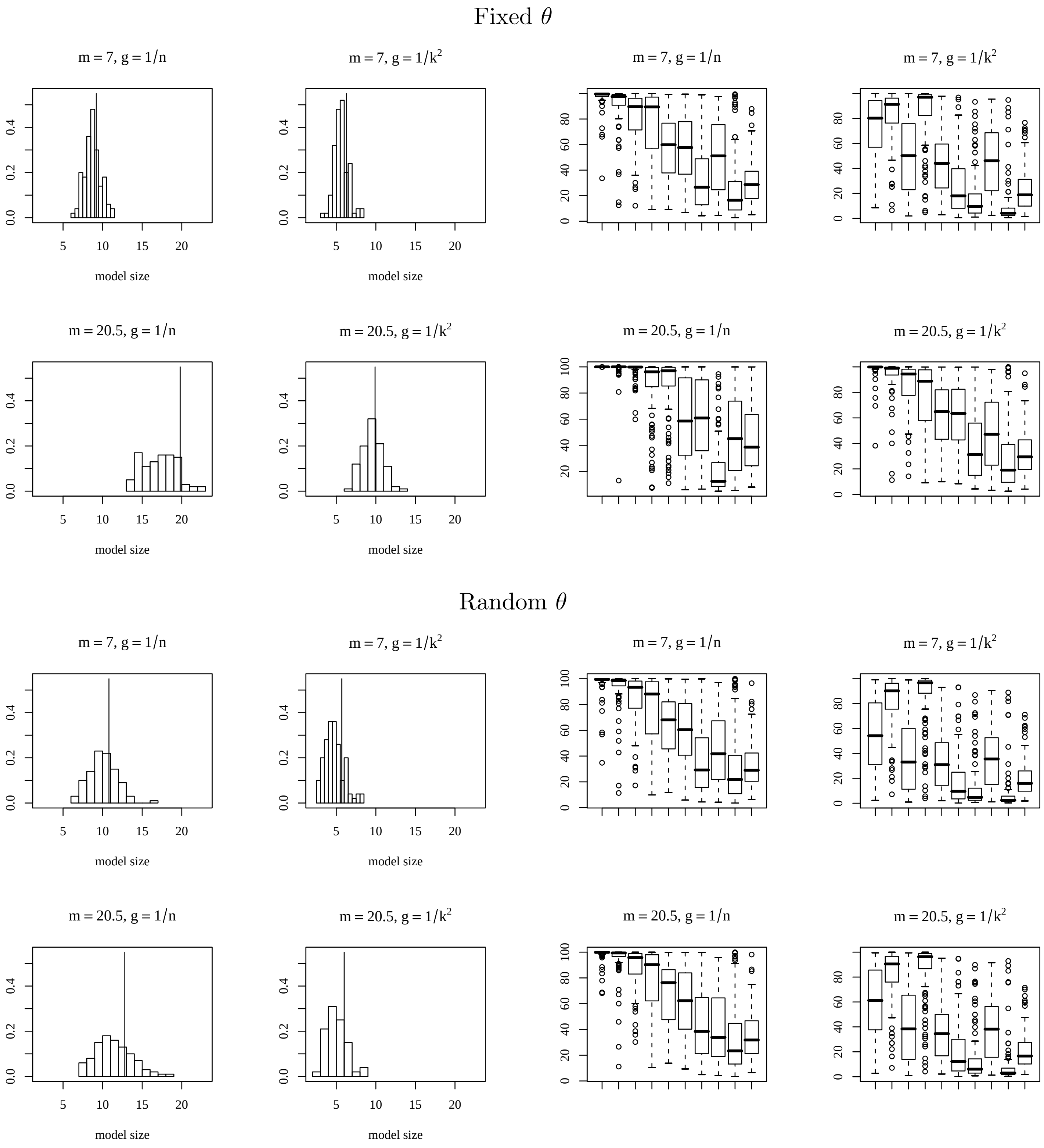

Fig. 6. Distribution of mean model size and posterior inclusion probabilities of the first ten regressors. (FLS data, 100 inference subsamples.)

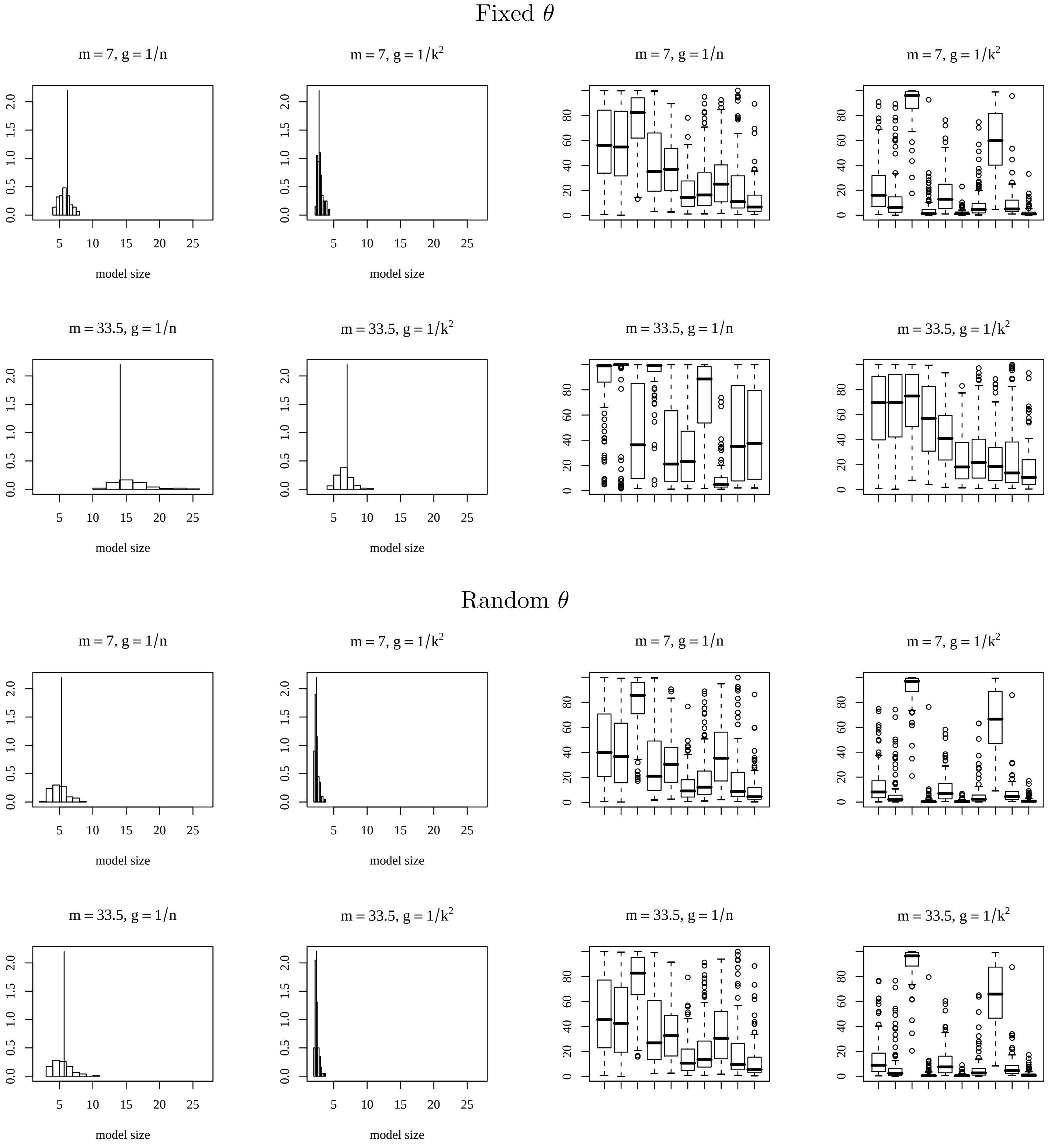

Fig. 7. Distribution of mean model size and posterior inclusion probabilities of the first ten regressors. (SDM data, 100 inference subsamples.)

6.2. Prediction

In order to compare the predictive performance associated with the various prior choices, we turn now to the task of predicting the observable, growth, given the regressors. Of course, the predictive distribution is also derived through model averaging, as explained in, *e.g.*, FLS. As a measure of how well each model predicts the retained observations, we use the log predictive score (LPS), which is a strictly proper scoring rule, described in FLS.[6] In the case of i.i.d. sampling, LPS can be given an interpretation in terms of the Kullback-Leibler divergence between the actual sampling density and the predictive density (Fernández *et al.*, 2001a), and smaller values indicate better prediction performance.

We compare the predictions based on: *(i)* BMA, *(ii)* the best model (the model with the highest posterior probability), *(iii)* the full model (with all k regressors), and *(iv)* the null model (with only the intercept).

Panel A in Table 5 summarizes our findings for the FLS data: the entries indicating "best" or "worst" model or how often a model is beaten by the null are expressed in percentages of the 100 samples for that particular prior setting. Selected runs with 500 subsamples lead to very similar results.

[1] *BMA*—The predictive performance of BMA is much superior to that of the other procedures—which corroborates evidence in *e.g.* Raftery *et al.*, 1997, Fernández *et al.*, 2001a and FLS. It is never the worst predictor and leads to the best predictions in more than half of the sampled cases (with the exception of the prior with fixed $\theta = 0.5$ and $g = 1/n$).

[2] *Best*—Basing predictions solely on the model with highest posterior probability is clearly a lot worse: it almost never gives the best prediction and leads to the worst prediction in 18 to 46% of the cases; moreover, it is beaten by the simple null model in more than 35% of the cases.

[3] *Full*—The use of the full model can lead to good forecasts, but is very risky, as it also has a substantial probability of delivering the worst performance: in fact, for $g = 1/k^2$ the proportion of the latter always exceeds the fraction of best performances. This behaviour is also illustrated by the fact that $\{\text{Min}, \text{Mean}, \text{Max}\}$ for LPS of the full model is $\{0.74, 1.77, 3.77\}$ for $g = 1/n$ and $\{0.67, 2.15, 5.82\}$ for $g = 1/k^2$. For comparison, the null model leads to $\{1.67, 2.05, 2.72\}$.

Having established that BMA is the strategy to adopt for prediction, we can focus on the forecast performance of BMA to compare the predictive ability across the different prior settings. Mean values of LPS (over all 100 samples) are not that different, but the maximum values indicate that fixing θ at 0.5 is the most risky strategy. This suggests using random θ. In addition, BMA always performs better with respect to the other prediction strategies for $g = 1/k^2$. Indeed, the worst case scenario for BMA appears to be $\theta = 0.5$ with $g = 1/n$. Finally, the choice of m almost leaves the random θ results unaffected, but has a substantial effect on the cases with fixed θ, in line with our expectations.

Panel B in Table 5 presents the same quantities for the MP data (where $k = 54$), and the $\{\text{Min}, \text{Mean}, \text{Max}\}$ values for LPS of the full model are $\{1.13, 2.55, 6.05\}$ for $g = 1/n$ and $\{1.12, 2.79, 7.48\}$ for $g = 1/k^2$, whereas the null model leads to $\{1.70, 2.03, 2.63\}$.

[6] Alternative scoring rules may be considered, like the continuous ranked probability score (CRPS). The CRPS measures the difference between the predicted and the observed cumulative distributions, and was found in *e.g.* Gneiting and Raftery (2007) to be less sensitive to outliers than LPS. It was introduced in the context of growth regressions by Eicher *et al.* (2007).

Table 5. Predictive performance: Three datasets (100 subsamples)

Panel A. FLS Data									
		$m = 7$				$m = k/2$			
		fixed θ		random θ		fixed θ		random θ	
		$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$	$g = 1/n$	$g = 1/k^2$
BMA	Best	59	65	54	66	48	63	54	67
(%)	Worst	0	0	0	0	0	0	0	0
	Beaten by Null	7	6	11	6	19	11	11	6
Best model	Best	1	2	1	0	3	2	2	0
(%)	Worst	28	18	34	18	46	31	36	17
	Beaten by null	35	35	41	36	52	47	45	36
Full model	Best	34	28	37	30	37	24	36	29
(%)	Worst	16	42	15	41	8	31	15	42
	Beaten by null	25	45	25	45	25	45	25	45
LPS of BMA	Minimum	1.12	1.19	1.12	1.21	0.86	1.11	1.11	1.20
	Mean	1.58	1.61	1.61	1.64	1.65	1.63	1.61	1.63
	Maximum	2.57	2.52	2.67	2.53	2.76	2.85	2.64	2.47
	St. dev.	0.30	0.26	0.33	0.25	0.42	0.37	0.34	0.25
Panel B. MP Data									
BMA	Best	71	70	75	73	61	78	73	72
(%)	Worst	0	0	0	0	0	0	0	1
	Beaten by null	15	16	15	18	32	15	16	18
Best model	Best	10	8	7	2	2	5	7	3
(%)	Worst	18	4	19	6	37	18	20	5
	Beaten by null	42	35	42	33	63	49	43	31
Full model	Best	4	7	3	9	6	2	4	9
(%)	Worst	61	75	59	75	46	66	59	75
	Beaten by null	71	77	71	77	71	77	71	77
LPS of BMA	Minimum	1.15	1.30	1.14	1.39	1.05	1.12	1.13	1.38
	Mean	1.70	1.74	1.71	1.77	1.87	1.72	1.72	1.77
	Maximum	3.34	3.46	3.26	3.53	3.60	3.31	3.23	3.54
	St. dev.	0.39	0.37	0.40	0.35	0.51	0.42	0.40	0.36
Panel C. SDM Data									
BMA	Best	69	53	64	48	48	63	66	46
(%)	Worst	0	0	0	0	0	0	0	0
	Beaten by null	15	11	16	10	51	16	16	10
Best model	Best	16	36	20	42	1	21	18	44
(%)	Worst	3	1	3	0	12	2	3	0
	Beaten by null	38	20	31	19	79	39	32	18
Full model	Best	0	0	0	0	0	0	0	0
(%)	Worst	97	99	97	100	88	98	97	100
	Beaten by null	99	100	99	100	99	100	99	100
LPS of BMA	Minimum	1.36	1.41	1.39	1.43	1.22	1.33	1.38	1.43
	Mean	1.78	1.75	1.78	1.75	2.18	1.81	1.79	1.74
	Maximum	3.38	2.67	3.31	2.53	3.70	3.63	3.38	2.51
	St. dev.	0.34	0.26	0.32	0.25	0.60	0.38	0.34	0.25

The superiority of BMA is even more pronounced, with the prior setting $\theta = 0.5$ with $g = 1/n$ again being the least favourable for BMA. The performance of the full model is now considerably worse than for the FLS data, as this model quite often leads to the worst behaviour and is soundly beaten by the much more conservative null model. This seems in line with the fact that the number of regressors, k, is now larger, so the full model is even more overparameterized than in the FLS case. Considering the LPS values for BMA, priors with $g = 1/n$ seem to have a slight edge, except for the case with a fixed $\theta = 0.5$. Finally, the choice of m is again virtually immaterial for the random θ cases, and affects those with fixed θ.

Finally, Panel C in Table 5 collects prediction results for the SDM data, where the number of regressors is even larger with $k = 67$. This is immediately felt in the behaviour of the full model, which is now virtually always the worst model. The $\{\text{Min}, \text{Mean}, \text{Max}\}$ values for LPS of the full model are $\{2.53, 5.16, 11.75\}$ for $g = 1/n$ and $\{2.65, 6.95, 23.21\}$ for $g = 1/k^2$, whereas the null model leads to $\{1.72, 2.08, 3.09\}$.

As before, the case with $\theta = 0.5$ and $g = 1/n$ leads to the worst BMA performance, allowing the null model to beat it more than half of the time! This is due to those cases where prediction is relatively difficult for which the conservative null model performs better. Throughout, the choice of $g = 1/k^2$ does better in avoiding large values of LPS, reducing the cases for which BMA is beaten by the null model. On the other hand, it is more conservative and lowers the percentage of cases where BMA leads to the best predictive behaviour. As we know from Subsection 4.3, the choice of $g = 1/k^2$ implies a larger size penalty, and thus stays closer to the rather conservative null model. As in the previous examples, the effect of m is far larger for fixed θ than for random θ. For the combinations of random θ with $g = 1/k^2$ BMA seems less dominant than for the other cases. This is simply a consequence of the fact that the best model then accounts for more than 60% of the probability mass (see Table 4), so that the best model also predicts very well here (albeit not quite as well as BMA). In fact, either BMA or the best model predict best in 90 of the 100 samples, while the null model does best for the remaining 10 samples.

If we measure robustness as in Subsection 6.1 by the standard deviation of the LPS values for BMA, Table 5 also shows us that cases with random θ and $g = 1/k^2$ are the most robust. Note that differences in robustness can be quite large, and the case with fixed $\theta = 0.5$ and $g = 1/n$ stands out as the least robust combination by far.

To summarize the predictive behaviour in terms of the choice of prior setting, we can conclude that a random θ prior seems preferable in view of the lack of sensitivity to m. Also, the combination of fixed $\theta = 0.5$ and $g = 1/n$ is to be avoided as it can lead to relatively bad forecasting behaviour and BMA is not as dominant as it is under other priors. For these other priors there seems no clear guidance for the choice of g on the basis of predictive behaviour alone.

7. Concluding Remarks and Recommendations

The theoretical and empirical evidence provided above shows the critical importance of prior assumptions for BMA: it clearly matters what prior settings we choose. The previously noted similarity of results with BACE and FLS prior settings turns out to be a fluke rather than an indication of prior robustness. Making certain prior assumptions implicit (as the choice of $g = 1/n$ in BACE) does not, in our view, constitute an improvement over a fully explicit Bayesian analysis and can easily lead to a false sense of prior robustness.

A first clear recommendation on the prior structure is to use random θ rather than fixed θ, since the hierarchical prior is much less sensitive to the (often rather arbitrary) choice of prior mean model size, m. Only in the unlikely situation when you really have very strong prior information on model size can a fixed θ prior be defensible. Therefore, we strongly discourage the use of the fixed θ prior as a "non-informative" prior, as it has clearly been shown to be quite informative.

Secondly, we would recommend to avoid the choice of $g = 1/n$, which implies a fairly small model size penalty and can, thus, result in convergence problems (with a very long tail of relatively unimportant models) and has also displayed more sensitivity to m than the alternative $g = 1/k^2$. In particular, we strongly advise against choosing $g = 1/n$ with fixed $\theta = 0.5$ as this combination can lead to relatively bad predictions and the superiority of BMA (which has been shown to be the best procedure to use for prediction) is less pronounced.

In conclusion: for growth-regression or other linear regression settings where we have a fairly large number of potential regressors with relatively few observations (where $k < n$ but of the same order of magnitude, so that $k^2 \gg n$), we would recommend to use the prior structure in (4) with $g = 1/k^2$ for any given model. Also, for the prior over models we strongly advise the use of the hierarchical prior on θ described in Subsection 4.1, whenever analysts have no really strong prior information on model size (which will typically be the case for growth regression). In that situation, the actual choice of the prior mean model size, m, will almost not matter, although we would, of course, advise to use a reasonable value for m.

8. References

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

Brock, W., and S. Durlauf (2001) “Growth Empirics and Reality,” *World Bank Economic Review*, 15: 229–72.

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.

Chipman, H., E.I. George and R.E. McCulloch (2001) “The Practical Implementation of Bayesian Model Selection,” (with discussion) in *Model Selection*, ed. P. Lahiri, IMS Lecture Notes, Vol. 38, pp. 70–134.

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

Doppelhofer, G., and M. Weeks (2008) “Jointness of Growth Determinants,” (with discussion) *Journal of Applied Econometrics*, forthcoming.

Durlauf, S.N., A. Kourtellos and C.M. Tan (2006) “Is God in the Details? A Reexamination of the Role of Religion in Economic Growth,” Economics Research Paper 2006-9, University of Wisconsin-Madison.

Eicher, T.S., C. Papageorgiou and A.E. Raftery (2007) “Determining Growth Determinants: Default Priors and Predictive Performance in Bayesian Model Averaging,” Working Paper No. 76, Center for Statistics and the Social Sciences, University of Washington, Seattle.

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.

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

George, E.I., and D.P. Foster (2000) “Calibration and Empirical Bayes variable selection,” *Biometrika*, 87: 731–747.

George, E.I., and R.E. McCulloch (1997) “Approaches for Bayesian Variable Selection,” *Statistica Sinica*, 7: 339–373.

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

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

Kass, R.E. and A.E. Raftery (1995), “Bayes factors” *Journal of the American Statistical Association* 90: 773–795.

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 (2005) “Mixtures of g-priors for Bayesian Variable Selection,” ISDS Discussion Paper 2005-12, Duke University.

León-González, R. and D. Montolio (2004) "Growth, Convergence and Public Investment: A BMA Approach," *Applied Economics*, 36: 1925–36.

Levine, R., and D. Renelt (1992) "A Sensitivity Analysis of Cross-Country Growth Regressions," *American Economic Review*, 82: 942–963.

Ley, E. and M.F.J. Steel (2007) "Jointness in Bayesian Variable Selection with Applications to Growth Regression," *Journal of Macroeconomics*, 29: 476–493.

Macleod, A.I. and D.R. Bellhouse (1983) "A Convenient Algorithm for Drawing a Simple Random Sample," *Applied Statistics*, 32: 182–184.

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

Masanjala, W. and C. Papageorgiou (2005) "Initial Conditions, European Colonialism and Africa's Growth," mimeo, Department of Economics, Louisiana State University, Baton Rouge.

Min, C.-K., and A. Zellner (1993) "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts With Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56: 89–118.

Raftery, A.E. (1995) "Bayesian Model Selection in Social Research," *Sociological Methodology*, 25: 111–163.

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. (1997) "I Just Ran Two Million Regressions," *American Economic Review*, 87: 178–183.

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.

Tsangarides, C.G. (2005) "Growth Empirics under Model Uncertainty: Is Africa Different?," IMF Working Paper 05/18, Washington, DC.

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.