Bayesian Model Averaging and Forecasting

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

Abstract. This paper focuses on the problem of variable selection in linear regression models. I briefly review the method of Bayesian model averaging, which has become an important tool in empirical settings with large numbers of potential regressors and relatively limited numbers of observations. Some of the literature is discussed with particular emphasis on forecasting in economics. The role of the prior assumptions in these procedures is highlighted, and some recommendations for applied users are given.

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

Address. M.F.J. Steel, Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K. Email: M.F.Steel@stats.warwick.ac.uk. Fax: +44(0)24 7652 4532. URL: http://www.warwick.ac.uk/go/msteel/

1. Introduction

Forecasting has a particularly critical role in economics and in many other sciences. However, forecasts are often very dependent on the model used, and the latter is not usually a given. There are many different models that researchers can use and the choice between them is far from obvious. Thus, the uncertainty over which model to use is an important aspect of forecasting or indeed any inference from data. Once we have realized that model uncertainty critically affects forecasts (and forecast uncertainty), the pooling of forecasts over models seems a natural way to deal with this. Such pooling can be done in many ways. The weights in combining different forecasts could simply be uniform or they could be based on historical performance (see for example Hendry and Clements, 2002 and Diebold and Lopez, 1996). A related methodology for dealing with large number of regressors used in macroeconomic forecasting is based on principal components or factors as in Stock and Watson (2002). Alternatively, the weights used in combining forecasts can be based on posterior model probabilities within a Bayesian framework. This procedure, which is typically referred to as Bayesian model averaging (BMA), is in fact the standard approach to model uncertainty within the Bayesian paradigm, where it is natural to reflect uncertainty through probability. Thus, it follows from direct application of Bayes' theorem as is explained in *e.g.* Leamer (1978), Min and Zellner (1993) and Raftery *et al.* (1997).

Min and Zellner (1993) show that mixing over models using BMA minimizes expected predictive squared error loss, provided the set of models under consideration is exhaustive. Raftery *et al.* (1997) state that BMA is optimal if predictive ability is measured by a logarithmic scoring rule. Thus, in the context of model uncertainty, the use of BMA follows from sensible utility considerations. In addition, empirical evidence of superior predictive performance of BMA can be found in, *e.g.*, Raftery *et al.* (1997), Fernández *et al.* (2001a,b) and Ley and Steel (2009). Predictive quality, in this literature, is typically measured in terms of a (proper) predictive score function that takes the entire predictive distribution into account. In particular, a popular quantity to use is the log predictive score. Alternative scoring rules are discussed in Gneiting and Raftery (2007).

I thank Toni Espasa for the kind invitation to contribute a paper to this special issue of the Bulletin. This paper was written while visiting the Department of Statistics at the Universidad Carlos III de Madrid, and I gratefully acknowledge their hospitality.

In economic applications we often face a large number of potential models with only a limited number of observations to conduct inference from. A good example is that of cross-country growth regressions. Insightful discussions of model uncertainty in growth regressions can be found in Brock and Durlauf (2001) and Brock *et al.* (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). Fernández *et al.* (2001b) 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. However, growth regressions is certainly not the only application area for BMA methods in economics. In the context of time series modelling, for example, it was used for posterior inference on impulse responses in Koop *et al.* (1994). Tobias and Li (2004) use BMA in the context of returns to education. The methodology has been applied by Garratt *et al.* (2003) for probability forecasting in the context of a small long-run structural vector error-correcting model of the U.K. economy. In addition, it was used for the evaluation of macroeconomic policy by Brock *et al.* (2003) and for the modelling of inflation by Cogley and Sargent (2005). Forecasting inflation using BMA has also been examined in Eklund and Karlsson (2007) and González (2010). Note that Eklund and Karlsson (2007) propose the use of so-called predictive weights in the model averaging, rather than the standard BMA, which is based on posterior model probabilities. An alternative strategy of dynamic model averaging, due to Raftery *et al.* (2010b) is used in Koop and Korobolis (2009) for inflation forecasting. The idea here is to use state-space models in order to allow for the forecasting model to change over time while, at the same time, allowing for coefficients in each model to evolve over time. Due to the use of approximations, the computations essentially boil down to the Kalman filter.

Even within the standard BMA context, there are different approaches in the literature. Differences between methods can typically be interpreted as the use of different prior assumptions. As shown in Ley and Steel (2009) and Eicher et al. (2011) prior assumptions can be extremely critical for the outcome of BMA analyses. For the use of BMA in practice it is important to understand how the (often almost arbitrarily chosen) prior assumptions

may affect our inference. The prior structures typically used in practice are designed to be "vague" choices, that require only a minimal amount of prior elicitation. More informative prior structures, such as the hierarchical model prior structures of Brock *et al.* (2003) can also be used, but of course, require more elicitation effort on the part of the user.

In 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 when we are dealing with comparing models (see, *e.g.*, Kass and Raftery, 1995). However, we should attempt to identify the effect of these assumptions in order to inform the analyst which prior settings are more informative than others, and in which direction they will influence the result. Ideally, prior structures are chosen so that they protect the user against unintended consequences of prior choices. For example, Ley and Steel (2009) advocate the use of hierarchical priors on model space, and show that this increases flexibility and decreases the dependence on essentially arbitrary prior assumptions.

For the sake of computational ease, sometimes approximations to the posterior model probabilities are used such as the BIC (Bayesian Information Criterion) used in Raftery (1995) and Sala-i-Martin *et al.* (2004). Ley and Steel (2009) show that the approach in Sala-i-Martin *et al.* (2004) corresponds quite closely to a BMA analysis with a particular choice of prior. It is important to stress that the dependence on prior assumptions does not disappear if we make those assumptions implicit rather than explicit. Thus, the claim in Sala-i-Martin *et al.* (2004) that their approach is less sensitive to prior assumptions is perhaps somewhat misleading.

For practitioners it is useful to know that freely available code for BMA exists on the web. In particular, there are R packages by Clyde (2010), Raftery *et al.* (2010a) and Feldkircher and Zeugner (2011). There is also code available in Fortran which is a development of the code in Fernández *et al.* (2001b) as explained in Ley and Steel (2007). The latter code is available at the *Journal of Applied Econometrics* data and code archive. Amini and Parmeter (2011) replicate some results in the literature with the R based packages and also explore frequentist model averaging techniques, such as the one in Magnus *et al.* (2010).

In this paper, we will, for simplicity, focus only on the application of standard BMA in the context of a linear regression model with uncertainty regarding the selection of explanatory variables. The next section briefly summarizes the main ideas of BMA. Section 3 describes the Bayesian model, and Section 4 examines some consequences of prior choices in more detail. The final section concludes.

2. The Principles of Bayesian Model Averaging

This section briefly presents the main ideas of BMA. When faced with 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 $M_j, j = 1, \ldots, J$, 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.

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.

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,

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

However, in practice J can be very large, making exhaustive computation of the sum in (1) prohibitively expensive in terms of computational effort. Then we often resort to simulation over $\mathcal{M}$. In particular, we can use a Markov chain Monte Carlo (MCMC) sampler to deal with the very large model space $\mathcal{M}$ (for example, containing 1.5×10^{20} models for the growth dataset of Sala-i-Martin *et al.* 2004 with 67 potential regressors). If the posterior odds between any two models are analytically available (as is the case for the model used in the next section), 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.

Other approaches to dealing with the large model space are the use of a coin-flip importance-sampling algorithm in Sala-i-Martin *et al.* (2004), and the branch-and-bound method developed by Raftery (1995). Eicher *et al.* (2011) 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. The Bayesian Adaptive Sampling approach of Clyde *et al.* (2011) adopts an alternative sampling strategy, which provides sequential learning of the marginal inclusion probabilities, while sampling without replacement. The latter is implemented in the R package of Clyde (2010). Bottolo and Richardson (2010) use Evolutionary Stochastic Search, which is designed to work for situations where the number of regressors is orders of magnitude larger than the number of observations (such as in genetic applications).

3. The Bayesian Model

Consider a Normal linear regression model for n observations of some response variable, 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. Model M_j will be the model with the $0 \leq k_j \leq k$ regressors grouped in Z_j, leading to

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

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.

Extensions to dynamic models are relatively straightforward, in principle, as some of the variables in Z can be lagged values of y. The use of BMA in panel data has been discussed in Moral-Benito (2010) in the context of growth regressions with country-specific fixed effects.

For the parameters in a given model M_j, Fernández *et al.* (2001a) propose 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

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

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. The prior on β_j in (4) should be proper, since 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 literature, (see, *e.g.*, Clyde and George, 2004).

Based on theoretical considerations and simulation results, Fernández *et al.* (2001a,b) choose to use $g = \max\{n, k^2\}$ in (4). In the literature, the two choices for g that underlie this recommendation are used quite frequently. The value $g = n$ 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. Choosing $g = k^2$ is suggested by the Risk Inflation Criterion of Foster and George (1994).

A natural Bayesian response to uncertainty (in this case, about which value of g to use) is to introduce a probability distribution. Thus, recent contributions suggest making g random by putting a hyperprior on g. In fact, the original Zellner-Siow prior can be

interpreted as such, and Liang *et al.* (2008) propose the class of hyper-g priors. Ley and Steel (2010) review the literature in this area and propose an alternative prior on g, which is compared with existing priors both in terms of theoretical properties (such as consistency) and in terms of empirical behaviour using various growth datasets and a dataset on returns to schooling (see Tobias and Li, 2004). Many of these prior distributions for g share the same form in that they correspond to a Beta prior for the shrinkage factor $g/(1+g)$.

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) and Fernández *et al.* (2001a,b) choose $\theta = 0.5$, which implies that $P(M_j) = 2^{-k}$ and that expected model size is $k/2$. Sala-i-Martin *et al.* (2004) 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.* (2011).

4. Prior Assumptions and Posterior Inference

4.1. Model prior specification and model size

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, 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 is typically done in most of the literature, 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 seems sensible. This idea was implemented by Brown *et al.* (1998), and is also discussed in, *e.g.*, Clyde and George (2004) and Nott and Kohn (2005). If we choose a Beta prior for θ with hyperparameters $a, b > 0$, *i.e.*, $\theta \sim \text{Be}(a, b)$, the prior mean model size is $\text{E}[W] = \frac{a}{a+b}k$. The implied prior distribution on model size is then a Binomial-Beta distribution (Bernardo and Smith, 1994, p. 117). In the case where $a = b = 1$ 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 Ley and Steel (2009) propose to facilitate prior elicitation by fixing $a = 1$. This still allows for a wide range of prior behaviour and makes it 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 $b = (k-m)/m$.

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 reflects a mildly conservative prior stance, where we require some data evidence to favour the inclusion of regressors.

For the case of $k = 67$ (corresponding to the growth data set in Sala-i-Martin et al., 2004), Figure 1 contrasts the prior model-size distributions with fixed θ (solid lines) and random θ (dashed lines), for two choices for mean model size: $m = 7$, which is used in Sala-i-Martin *et al.* (2004), 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 quite concentrated. Treating θ as random will typically imply more prior uncertainty about model size, which is often more reasonable in practice.

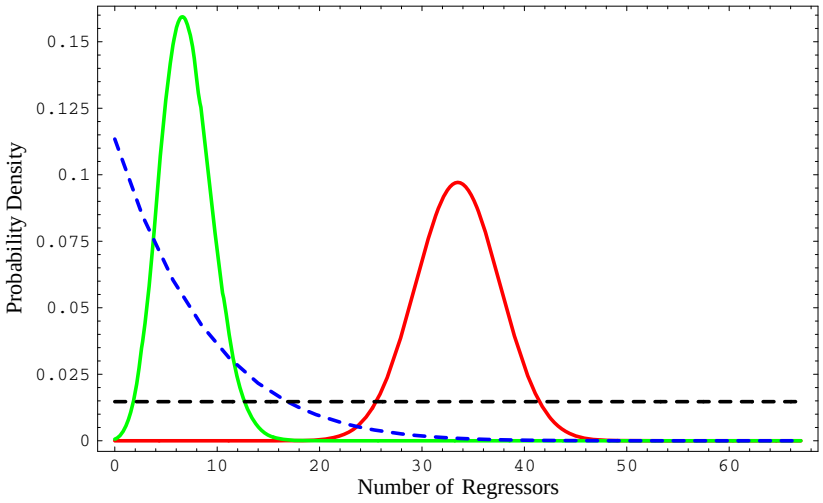

Fig. 1. Prior model size for $k = 67$; fixed θ (solid) and random θ (dashed).

For ease of presentation, these discrete distributions are depicted through continuous graphs.

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). Generally, if we fix θ and express things in terms of the prior mean model size m, these prior odds are

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

from which it is clear that the prior favours larger models if $m > k/2$. For the hierarchical Be(a, b) prior on θ, with $a = 1$ and the prior elicitation in terms of m we obtain the prior odds:

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

Ley and Steel (2009) illustrate 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 of size around $k/2$ in the model space $\mathcal{M}$ than of size nearer to 0 or k. 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 $m < k/2$.

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 illustrated by Figure 2, 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.

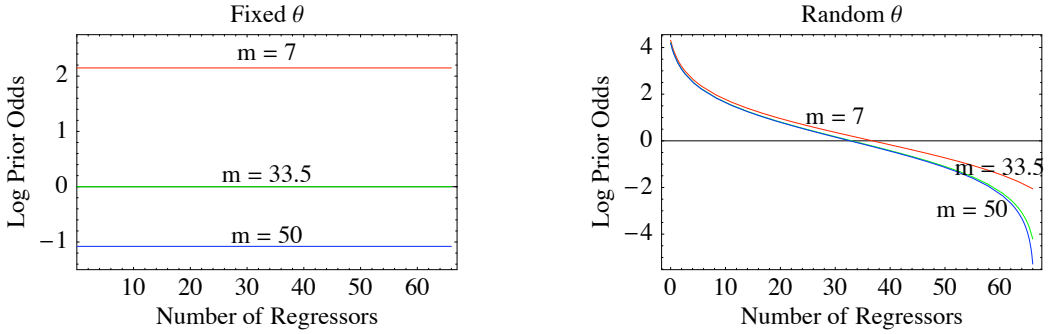

Fig. 2. Log of Prior Odds: $k_i = (k_j - 1)$ vs varying k_j.

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 less robust 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. If g in (4) is fixed and does not depend on the model size k_j, the Bayes factor for any two models from (3)–(4) becomes:

$$\frac{l_y(M_i)}{l_y(M_j)} = (1+g)^{\frac{k_j-k_i}{2}} \left(\frac{1+g(1-R_i^2)}{1+g(1-R_j^2)} \right)^{-\frac{n-1}{2}}, \tag{5}$$

where R_i^2 is the usual coefficient of determination for model M_i. The expression in (5) 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. Ley and Steel (2009) remark that this expression is very close to the one used in Sala-i-Martin *et al.* (2004) provided we take $g = n$.

If we assign a hyperprior to g, then we need to deal with the fact that, generally, the integral of (5) with respect to g does not have a straightforward closed-form solution. Liang *et al.* (2008) approximate this integral with a Laplace approximation, but Ley and Steel (2009) formally integrate out (5) with g and implement this with a Gibbs sampler approach over model space and g. In the latter, the Bayes factor between any two models given g is given by (5), and the conditional posterior of g given M_j is

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

where $p(g \mid M_j)$ is the hyperprior on g (which could potentially depend on M_j). The advantage of conducting posterior inference on (g, M_j) is that it does not rely on approximations and it 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.

It is interesting to examine more in detail how the various prior choices translate into model size penalties. In the case of fixed g we deduce from (5) that if two models fit equally well (*i.e.*, $R_i^2 = R_j^2$), then the Bayes factor will approximately equal $g^{(k_j - k_i)/2}$. If one of the models contains one more regressor, this means that the larger model will be penalized by $g^{-1/2}$. Thus, both the choices of g and m have an implied model size penalty. Ley and Steel (2009) and Eicher *et al.* (2011) examine the trade-off between choosing m and g in this context.

Ley and Steel (2010) compare various hyperpriors on g and conclude that the hyper-g/n prior of Liang *et al.* (2008) and the benchmark prior they propose are fairly safe choices for use in typical economic applications.

5. Concluding Remarks

The use of Bayesian Model Averaging is rapidly becoming an indispensable tool in economics to deal with model uncertainty. Whereas this is a powerful tool with a strong foundation in statistical theory, the empirical results of such procedures can, however, be quite sensitive to prior assumptions. It is therefore important that we investigate the effect of various prior structures in order to be able to give sensible recommendations to practitioners. The use of hierarchical priors seems a good way to specify robust priors, and, in the context of variable selection in the linear regression model with a number of observations of the same order of magnitude as the number of potential covariates, I would recommend the use of a hierarchical prior on θ (with a sensible value for prior mean model size) and the use of a hyperprior on g of the types mentioned at the end of the previous section.

6. References

Amini, S. and C. Parmeter (2011) "Comparison of Model Averaging Techniques: Assessing Growth Determinants" mimeo, Virginia Polytechnic Institute and State University.

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

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

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. (2010), *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging*. R package vs. 0.92. URL: `http://CRAN.R-project.org/package=BAS`

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

Clyde, M.A., J. Ghosh, M. L. Littman (2011) "Bayesian Adaptive Sampling for Variable Selection and Model Averaging," *Journal of Computational and Graphical Statistics*, 20: 80–101.

Cogley T and T.J. Sargent (2005) "The Conquest of US Inflation: Learning and Robustness to Model Uncertainty," *Review of Economic Dynamics*, 8: 528-563.

Diebold F.X. and J.A. Lopez (1996) "Forecast Evaluation and Combination". In Handbook of Statistics, Maddala GS, Rao CR (eds.); North-Holland: Amsterdam

Durlauf, S.N., A. Kourtellos and C.M. Tan (2011) "Is God in the Details? A Reexamination of the Role of Religion in Economic Growth," *Journal of Applied Econometrics*, forthcoming.

Eklund, J. and S. Karlsson (2007) "Forecast Combination and Model Averaging Using Predictive Measures," *Econometric Reviews*, 26: 329–363.

Eicher, T.S., C. Papageorgiou and A.E. Raftery (2011) "Default Priors and Predictive Performance in Bayesian Model Averaging, with Application to Growth Determinants, *Journal of Applied Econometrics*, 26: 30-55.

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.

Garratt A, K. Lee, M.H. Pesaran and Y. Shin (2003) "Forecasting Uncertainties in Macroeconometric Modelling: An Application to the UK Economy," *Journal of the American Statistical Association*, 98: 829-838.

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

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

González, E. (2010) "Bayesian Model Averaging. An Application to Forecast Inflation in Colombia," Borradores de Economía 604, Banco de la República, Colombia.

Hendry, D.F. and M.P. Clements (2002) "Pooling of Forecasts," *Econometrics Journal*, 5: 1-26.

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.

Koop G. and D. Korobolis (2009) "Forecasting Inflation Using Dynamic Model Averaging," Technical Report 34-09, Rimini Center for Economic Analysis.

Koop, G., J. Osiewalski and M.F.J. Steel (1994) "Posterior Properties of Long-run Impulse Responses," *Journal of Business and Economic Statistics* 12: 489–492.

Leamer, E.E., 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley, New York.

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.

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.

Ley, E. and M.F.J. Steel (2010) "Mixtures of g-priors for Bayesian Model Averaging with Economic Applications," CRiSM Working Paper 10-23, University of Warwick.

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.

Magnus, J.R., Powell, O. and P. Prüfer (2010) "A Comparison of Two Model Averaging Techniques with an Application to Growth Empirics," *Journal of Econometrics*, 154: 139–153.

Moral-Benito, E. (2010) "Determinants of Economic Growth: A Bayesian Panel Data Approach," mimeo, CEMFI.

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.

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

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

Raftery, A.E., J. A. Hoeting, C. Volinsky, I. Painter and K.Y. Yeung (2010a), *BMA: Bayesian Model Averaging.* R package vs. 3.13. URL: `http://CRAN.R-project.org/package=BMA`

Raftery, A.E., M. Kárný, and P. Ettler (2010b) "Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill," *Technometrics*, 52: 52-66.

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.

Stock, J. and Watson, M. (2002) "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97: 1167-1179.

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.

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.

Feldkircher, M. and S. Zeugner (2011) "Bayesian Model Averaging with BMS," R-package vs. 0.3.0, URL: `http://cran.r-project.org/web/packages/BMS/`