Model Averaging and its Use in Economics

Mark F.J. Steel *
Department of Statistics, University of Warwick

September 22, 2017

Summary

The method of model averaging has become an important tool to deal with model uncertainty, in particular in empirical settings with large numbers of potential models and relatively limited numbers of observations, as are common in economics. Model averaging is a natural response to model uncertainty in a Bayesian framework, so most of the paper deals with Bayesian model averaging. In addition, frequentist model averaging methods are also discussed. Numerical methods to implement these methods are explained, and I point the reader to some freely available computational resources. The main focus is on the problem of variable selection in linear regression models, but the paper also discusses other, more challenging, settings. Some of the applied literature is reviewed with particular emphasis on applications in economics. The role of the prior assumptions in Bayesian procedures is highlighted, and some recommendations for applied users are provided (JEL: C11, C15, C20, C52, O47).

1 Introduction

This paper is about model averaging, as a solution to the problem of model uncertainty and focuses mostly on the theoretical developments over the last two decades and its uses in applications in economics. This is a topic that has now gained substantial maturity and is generating a rapidly growing literature. Thus, a survey seems timely. The discussion focuses mostly on covariate selection in regression models (normal linear regression and its extensions), which is arguably the most pervasive situation in economics. Advances in the context of models designed

*Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK; email: m.steel@warwick.ac.uk. I am grateful to the Editor for giving me the opportunity to write this article and for stimulating comments. Insightful comments from Jesús Crespo Cuaresma, Anabel Forte, Gonzalo García-Donato, Anne Miloschewski, Chris Papageorgiou, David Rossell and Hal Varian were very gratefully received and used to improve the paper. In July 2013, Eduardo Ley, who has made substantial contributions in this area, tragically passed away. He was a very dear friend and a much valued coauthor and this paper is dedicated to his memory.

to deal with some more challenging situations, such as data with dependency over time or in space or endogeneity (all quite relevant in economic applications) are also discussed. Two main strands of model averaging are distinguished: Bayesian model averaging (BMA), based on probability calculus and naturally emanating from the Bayesian paradigm by treating the model index as an unknown, just like the model parameters and specifying a prior on both; and frequentist model averaging (FMA), where the chosen weights are often determined so as to obtain desirable properties of the resulting estimators under repeated sampling and asymptotic optimality.

In particular, the aim of this paper is threefold:

- To provide a survey of the most important methodological contributions in a consistent notation and through a formal, yet accessible, presentation. This review takes into account the latest developments and applications, which is important in such a rapidly developing literature. Technicalities are not avoided, but some are dealt with by providing the interested reader with the relevant references. Even though the list of references is quite extensive, this is not claimed to be an exhaustive survey. Rather, it attempts to identify the most important developments that the applied economist needs to know about for an informed use of these methods. This review complements and extends other reviews and discussions; for example by Hoeting et al. (1999) on Bayesian model averaging, Clyde and George (2004) on model uncertainty, Moral-Benito (2015) on model averaging in economics and Wang et al. (2009) on frequentist model averaging. Further, we can mention a review of weighted average least squares in Magnus and De Luca (2016) while Fragoso and Neto (2015) develop a conceptual classification scheme to better describe the literature in Bayesian model averaging. Koop (2017) discusses the use of Bayesian model averaging or prior shrinkage as responses to the challenges posed by big data in empirical macroeconomics.

- By connecting various strands of the literature, to enhance the insight of the reader into the way these methods work and why we would use them. In particular, this paper attempts to tie together disparate literatures with roots in econometrics and statistics, such as the literature on forecasting, often in the context of time series and linked with information criteria, fundamental methodology to deal with model uncertainty and shrinkage in statistics[1], as well as more ad-hoc ways of dealing with variable selection. There is also a subsection explaining the various commonly used numerical methods to implement model averaging

[1] Variable selection can be interpreted as a search for parsimony, which has two main approaches in Bayesian statistics: through the use of shrinkage priors, which are absolutely continuous priors that shrink coefficients to zero but where all covariates are always included in the model, and through allocating prior point mass at zero for each of the regression coefficients, which allows for formal exclusion of covariates and implies that we need to deal with many different models, which is the approach recommended here.

in practical situations, which are typically characterized by very large model spaces. For Bayesian model averaging, it is important to understand that the weights (based on posterior model probabilities) are typically quite sensitive to the prior assumptions, in contrast to the usually much more robust results for the model parameters given a specific model. In addition, this sensitivity does not vanish as the sample size grows (Kass and Raftery, 1995; Berger and Pericchi, 2001). Thus, a good understanding of the effect of (seemingly arbitrary) prior choices is critical.

- To provide sensible recommendations for empirical researchers about which modelling framework to adopt and how to implement these methods in their own research. In the case of Bayesian model averaging, I recommend the use of prior structures that are easy to elicit and are naturally robust. I include a separate section on freely available computational resources that will allow applied researchers to try out these methods on their own data, without having to incur a prohibitively large investment in implementation. In making recommendations, it is inevitable that one draws upon personal experiences and preferences, to some extent. Thus, I present the reader with a somewhat subjective point of view, which I believe, however, is well-supported by both theoretical and empirical results.

Given the large literature, and in order to preserve a clear focus, it is important to set some limits to the coverage of the paper. As already explained above, the paper deals mostly with covariate selection in regression models, and does not address issues like the use of Bayesian model averaging in classification trees (Hernández et al., 2017) or in clustering and density estimation (Russell et al., 2015). The large literature in machine learning related to nonparametric approaches to covariate selection (Hastie et al., 2009) will also largely be ignored. Finally, this paper considers situations where the number of observations exceeds the number of potential covariates as this is most common in economics (some brief comments on the opposite case can be found in footnote 10).

As mentioned above, I discuss Bayesian and frequentist approaches to model averaging. This paper is mostly concerned with the Bayesian approach for two reasons:

- I personally find the formality and probability-based interpretability of the Bayesian approach very appealing, as opposed to the more ad-hoc nature and often asymptotic motivation of the frequentist methodology. I do realize this is, to some extent, a personal choice, but I prefer to operate within a logically closed and well-defined methodological framework, which immediately links to prediction and decision theory and where it is made explicit what the user adds to the analysis (in particular, the choice of priors).

- There is a large amount of recent literature on the Bayesian approach to resolving model uncertainty, both in statistics and in many areas of application, among which economics features rather prominently. Thus, this focus on Bayesian methods is in line with the majority of the literature (see footnote 5) and seems to reflect the perceived preference of many researchers in economics.

Of course, as Wright (2008) states: "One does not have to be a subjectivist Bayesian to believe in the usefulness of BMA, or of Bayesian shrinkage techniques more generally. A frequentist econometrician can interpret these methods as pragmatic devices that may be useful for out-of-sample forecasting in the face of model and parameter uncertainty."

This paper is organised as follows: in Section 2 I discuss the issue of model uncertainty and the way it can be addressed through Bayesian model averaging, and I introduce the specific context of covariate selection in the normal linear model. Section 3 provides a detailed account of Bayesian model averaging, focusing on the prior specification, its properties and its implementation in practice. This section also provides a discussion of various generalizations of the sampling model and of a number of more challenging models, such as dynamic models and models with endogenous covariates. Section 4 describes frequentist model averaging, while Section 5 surveys some of the recent literature where model averaging methods have been applied in economics. In Section 6 some freely available computational resources are briefly discussed, and the final section concludes and provides some recommendations.

2 Model uncertainty

The issue of model uncertainty is a very important one, particularly for modelling in the social sciences, where there is usually a large amount of uncertainty about model specifications that is not resolved by universally accepted theory. Thus, it affects virtually all modelling in economics and its consequences need to be taken into account whenever we are (as usual) interested in quantities that are not model-specific (such as predictions or effects of regressors). Generally, one important and potentially dangerous consequence of neglecting model uncertainty is that we assign more precision to our inference than is warranted by the data, and this leads to overly confident decisions and predictions. In addition, our inference can be severely biased. See Chatfield (1995) and Draper (1995) for extensive discussions of model uncertainty. In the context of the evaluation of macroeconomic policy, Brock et al. (2003) describe and analyse some approaches to dealing with the presence of uncertainty about the structure of the economic environment under study. Starting from a decision-theoretic framework, they recommend model averaging as a critical tool in tackling uncertainty. Brock and Durlauf (2015) specifically focus on policy evaluation and

provide an overview of different approaches, distinguishing between cases in which the analyst can and cannot provide conditional probabilities for the effects of policies. Varian (2014) states that "An important insight from machine learning is that averaging over many small models tends to give better out-of-sample prediction than choosing a single model. In 2006, Netflix offered a million dollar prize to researchers who could provide the largest improvement to their existing movie recommendation system. The winning submission involved a 'complex blending of no fewer than 800 models,' though they also point out that 'predictions of good quality can usually be obtained by combining a small number of judiciously chosen methods' (Feuerverger et al., 2012). It also turned out that a blend of the best- and second-best submissions outperformed either of them. Ironically, it was recognized many years ago that averages of macroeconomic model forecasts outperformed individual models, but somehow this idea was rarely exploited in traditional econometrics. The exception is the literature on Bayesian model averaging, which has seen a steady flow of work; see Steel (2011) for a survey."

As an example, Durlauf et al. (2012) examine the effect of different substantive assumptions about the homicide process on estimates of the deterrence effect of capital punishment[2]. Considering four different types of model uncertainty, they find a very large spread of effects, with the estimate of net lives saved per execution ranging from -63.6 (so no deterrence effect at all) to 20.9. This clearly illustrates that the issue of model uncertainty needs to be addressed before we can answer questions such as this and many others of immediate relevance to society.

Over the last decade, there has been a rapidly growing awareness of the importance of dealing with model uncertainty for economics. As examples, the *European Economic Review* has recently published a special issue on "Model Uncertainty in Economics" which was also the subject of the 2014 Schumpeter lecture in Marinacci (2015), providing a decision-theory perspective. In addition, a book written by two Nobel laureates in economics (Hansen and Sargent, 2014), focuses specifically on the effects of model uncertainty on rational expectations equilibrium concepts.

In line with probability theory, the standard Bayesian response to dealing with uncertainty is to average. When dealing with parameter uncertainty, this involves averaging over parameter values with the posterior distribution of that parameter in order to get the predictive distribution. Analogously, model uncertainty is also resolved through averaging, but this time averaging over models with the (discrete) posterior model distribution. The latter procedure is usually called Bayesian model averaging and was already described in Leamer (1978) and later used in Min and Zellner (1993), Koop et al. (1997) and Raftery et al. (1997a). BMA thus appears as a direct consequence of Bayes theorem (and hence probability laws) in a model uncertainty setting and is perhaps best introduced by considering the concept of a predictive distribution, often of interest

[2] A systematic investigation of this issue goes back to Leamer (1983).

in its own right. In particular, assume we are interested in predicting the unobserved quantity y_f on the basis of the observations y. Let us denote the sampling model[3] for y_f and y jointly by $p(y_f|y, \theta_j, M_j)p(y|\theta_j, M_j)$, where M_j is the model selected from a set of K possible models, and $\theta_j \in \Theta_j$ groups the (unknown) parameters of M_j. In a Bayesian framework, any uncertainty is reflected by a probability distribution[4] so we assign a (typically continuous) prior $p(\theta_j|M_j)$ for the parameters and a discrete prior $P(M_j)$ defined on the model space. We then have all the building blocks to compute the predictive distribution as

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

where the quantity in square brackets is the predictive distribution given M_j obtained using the posterior of θ_j given M_j, which is computed as

$$p(\theta_j|y, M_j) = \frac{p(y|\theta_j, M_j)p(\theta_j|M_j)}{\int_{\Theta_j} p(y|\theta_j, M_j)p(\theta_j|M_j)\mathrm{d}\theta_j} \equiv \frac{p(y|\theta_j, M_j)p(\theta_j|M_j)}{p(y|M_j)}, \tag{2}$$

with the second equality defining $p(y|M_j)$, which is used in computing the posterior probability assigned to M_j as follows:

$$P(M_j|y) = \frac{p(y|M_j)P(M_j)}{\sum_{i=1}^{K} p(y|M_i)P(M_i)} \equiv \frac{p(y|M_j)P(M_j)}{p(y)}. \tag{3}$$

Clearly, the predictive in (1) indeed involves averaging at two levels: over (continuous) parameter values, given each possible model, and discrete averaging over all possible models. The denominators of both averaging operations are not immediately obvious from (1), but are made explicit in (2) and (3). The denominator (or integrating constant) $p(y|M_j)$ in (2) is the so-called marginal likelihood of M_j and is a key quantity for model comparison. In particular, the Bayes factor between two models is the ratio of their marginal likelihoods and the posterior odds are directly obtained as the product of the Bayes factor and the prior odds. The denominator in (3), $p(y)$, is defined as a sum and the challenge in its calculation often lies in the sheer number of possible models, i.e. K.

Bayesian model averaging as described above is thus the formal probabilistic way of obtaining predictive inference, and is, more generally, the approach to any inference problem involving quantities of interest that are not model-specific. So it is also the Bayesian solution to conducting posterior inference on *e.g.* the effects of covariates. Formally, the posterior distribution of any

[3]For ease of notation, we will assume continuous sampling models with real-valued parameters throughout, but this can immediately be extended to other cases.

[4]Or, more generally, a measure.

quantity of interest, say Δ, which has a common interpretation across models is a mixture of the model-specific posteriors with the posterior model probabilities as weights, *i.e.*

$$P_{\Delta|y} = \sum_{j=1}^{K} P_{\Delta\,|\,y,M_j} P(M_j\,|\,y). \tag{4}$$

The rapidly growing importance of model averaging as a solution to model uncertainty is illustrated by Figure 1, which plots the citation profile over time of papers with the topic "model averaging" in the literature (left panel). A large part of the literature uses Bayesian model averaging methods, reflected in the citations to papers with the topic "Bayesian model averaging" shown in the right panel of Figure 1[5]. The sheer number of recent papers in this area is evidenced by the fact that Google Scholar returns over 45,000 papers in a search for "model averaging" and 17,000 papers when searching for "Bayesian model averaging", over half of which date from 2012 and later (data from September 7, 2017).

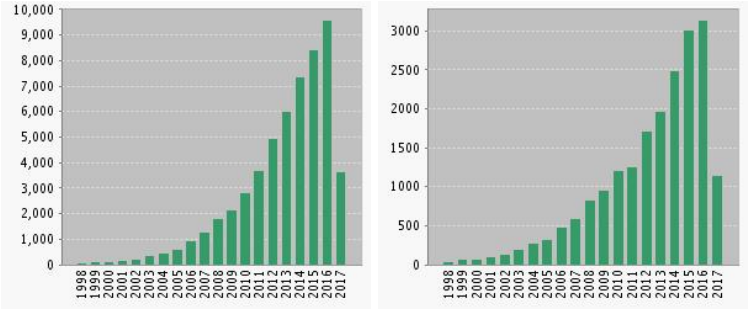

Figure 1: Left: number of citations to papers with topic "model averaging". Right: number of citations to papers with topic "Bayesian model averaging". Source: Web of Science, June 4, 2017

2.1 Covariate selection in the normal linear regression model

Most of the relevant literature assumes the simple case of the normal linear sampling model. This helps tractability, and is fortunately also a model that is often used in empirical work. We shall follow this tradition, and will assume for most of the paper[6] that the sampling model is normal

[5]Comparison of both graphs in Figure 1 might create the mistaken impression that most of the literature uses frequentist rather than Bayesian model averaging methods. However, the equivalent search for "frequentist model averaging" only generates less than 1% of the total citations represented in the right hand panel of the figure.

[6]Section 3.8 explores some extensions, *e.g.* to the wider class of Generalized Linear Models (GLMs) and some other modelling frameworks that deal with specific challenges in economics.

with a mean which is a linear function of some covariates[7]. We shall further assume, again in line with the vast majority of the literature (and many real-world applications) that the model uncertainty relates to the choice of which covariates should be included in the model, *i.e.* under model j the n observations in y are generated from

$$y|\theta_j, M_j \sim N(\alpha\iota + Z_j\beta_j, \sigma^2). \tag{5}$$

Here ι represents a $n \times 1$-dimensional vector of ones, Z_j groups k_j of the possible k regressors (*i.e.* it selects k_j columns from an $n \times k$ matrix Z, corresponding to the full model) and $\beta_j \in \Re^{k_j}$ are its corresponding regression coefficients. Furthermore, all considered models contain an intercept $\alpha \in \Re$ and the scale $\sigma > 0$ has a common interpretation across all models. We standardize the regressors by subtracting their means, which makes them orthogonal to the intercept and renders the interpretation of the intercept common to all models. The model space is then formed by all possible subsets of the covariates and thus contains $K = 2^k$ models in total[8]. Therefore, the model space includes the null model (the model with only the intercept and $k_j = 0$) and the full model (the model where $Z_j = Z$ and $k_j = k$). This definition of the model space is consistent with the typical situation in economics, where theories regarding variable inclusion do not necessarily exclude each other. Brock and Durlauf (2001) refer to this as the "open-endedness" of the theory[9]. Throughout, we assume that the matrix formed by adding a column of ones to Z has full column rank[10].

This model uncertainty problem is very relevant for empirical work, especially in the social sciences where typically competing theories abound on which are the important determinants of a phenomenon. Thus, the issue has received quite a lot of attention both in statistics and economics, and various approaches have been suggested. We can mention:

[7]Note that this is not as restrictive as it seems. It certainly does not mean that the effects of determinants on the modelled phenomenon are linear; we can simply include regressors that are nonlinear transformations of determinants, interactions etc.

[8]This can straightforwardly be changed to a (smaller) model space where some of the regressors are always included in the models.

[9]In the context of growth theory, Brock and Durlauf (2001) define this concept as "the idea that the validity of one causal theory of growth does not imply the falsity of another. So, for example, a causal relationship between inequality and growth has no implications for whether a causal relationship exists between trade policy and growth."

[10]For economic applications this is generally a reasonable assumption, as typically $n > k$, although they may be of similar orders of magnitude. In other areas such as genetics this is usually not an assumption we can make. However, it generally is enough that for each model we consider to be a serious contender the matrix formed by adding a column of ones to Z_j is of full column rank, and that is much easier to ensure. Implicitly, in such situations we would assign zero prior and posterior probability to models for which $k_j \geq n$. Formal approaches to use g-priors in situations where $k > n$ include Maruyama and George (2011) and Berger et al. (2016), based on different ways of generalizing the notion of inverse matrices.

1. Stepwise regression: this is a sequential procedure for entering and deleting variables in a regression model based on some measure of "importance", such as the t-statistics of their estimated coefficients (typically in "backwards" selection where covariates are considered for deletion) or (adjusted) R^2 (typically in "forward" selection when candidates for inclusion are evaluated).

2. Shrinkage methods: generally, these methods aim to find a set of sparse solutions (i.e. models with a reduced set of covariates) by shrinking coefficient estimates toward zero. Bayesian shrinkage methods rely on the use of shrinkage priors, which are such that some of the estimated regression coefficients in the full model will be close to zero. Typically, such priors will have a peak at zero to induce shrinkage for small coefficients, combined with fat tails that do not lead to shrinkage for large coefficients. Examples are the normal-gamma prior of Griffin and Brown (2010) and the horseshoe prior of Carvalho et al. (2010). A common classical method is penalized least squares, such as LASSO (least absolute shrinkage and selection operator), introduced by Tibshirani (1996), where the regression "fit" is maximized subject to a complexity penalty. Choosing a different penalty function, Fan and Li (2001) propose the smoothly clipped absolute deviation (SCAD) penalized regression estimator.

3. Information criteria: these criteria can be viewed as the use of the classical likelihood ratio principle combined with penalized likelihood (where the penalty function depends on the model complexity). A common example is the Akaike information criterion (AIC). The Bayesian information criterion (BIC) implies a stronger complexity penalty and was originally motivated through asymptotic equivalence with a Bayes factor (Schwarz, 1978). Asymptotically, AIC selects a single model that minimizes the mean squared error of prediction. BIC, on the other hand, chooses the correct model with probability tending to one as the sample size grows to infinity. So BIC is consistent, while AIC is not. Spiegelhalter et al. (2002) propose the Deviance information criterion (DIC) which can be interpreted as a Bayesian generalization of AIC.[11]

4. Cross-validation: the idea here is to use only part of the data for inference and to assess how well the remaining observations are predicted by the fitted model. This can be done repeatedly for random splits of the data and models can be chosen on the basis of their predictive performance.

5. Extreme Bounds Analysis (EBA): this procedure was proposed in Leamer (1983, 1985)

[11] DIC is quite easy to compute in practice, but has been criticized for its dependence on the parameterization and its lack of consistency.

and is based on distinguishing between "core" and "doubtful" variables. Rather than a discrete search over models that include or exclude subsets of the variables, this sensitivity analysis answers the question: how extreme can the estimates be if any linear homogenous restrictions on a selected subset of the coefficients (corresponding to doubtful covariates) are allowed? An extreme bounds analysis chooses the linear combinations of doubtful variables that, when included along with the core variables, produce the most extreme estimates for the coefficient on a selected core variable. If the extreme bounds interval is small enough to be useful, the coefficient of the core variable is reported to be "sturdy".

6. s-values: proposed by Leamer (2016a,b) as a measure of "model ambiguity". Here σ is replaced by the OLS estimate and no prior mass points at zero are assumed for the regression coefficients. For each coefficient, this approach finds the interval bounded by the extreme estimates (based on different prior variances, elicited through R^2); the s-value (s for sturdy) then summarizes this interval of estimates in the same way that a t-statistic summarizes a confidence interval (it simply reports the centre of the interval divided by half its width). A small s-value then indicates fragility of the effect of the associated covariate, by measuring the extent to which the sign of the estimate of a regression coefficient depends on the choice of model.

7. General-to-specific modelling: this approach starts from a general unrestricted model and uses individual t-statistics to reduce the model to a parsimonious representation. We refer the reader to Hoover and Perez (1999) and Hendry and Krolzig (2005) for background. Hendry and Krolzig (2004) present an application of this technique to the cross-country growth dataset of Fernández et al. (2001b) ("the FLS data", which record average per capita GDP growth over 1960-1992 for $n = 72$ countries with $k = 41$ potential regressors).

8. The Model Confidence Set (MCS): this approach to model uncertainty consists in constructing a set of models such that it will contain the best model with a given level of confidence. This was introduced by Hansen et al. (2011) and only requires the specification of a collection of competing objects (model space) and a criterion for evaluating these objects empirically. The MCS is constructed through a sequential testing procedure, where an equivalence test determines whether all objects in the current set are equally good. If not, then an elimination rule is used to delete an underperforming object. The same significance level is used in all tests, which allows one to control the p-value of the resulting set and each of its elements. The appropriate critical values of the tests are determined by bootstrap procedures. Hansen et al. (2011) apply their procedure to e.g. US inflation forecasting, and Wei and Cao (2017) use it for modelling Chinese house prices, using predictive elimination criteria.

9. Best subset regression of Hastie et al. (2009), called full subset regression in Hanck (2016). This method considers all possible models: for a given model size k_j it selects the best in terms of fit (the lowest sum of squared residuals). As all these models have k_j parameters, none has an unfair advantage over the others using this criterion. Of the resulting set of optimal models of a given dimension, the procedure then chooses the one with the smallest value of some criterion such as Mallows' C_p [12]. Hanck (2016) does a small simulation exercise to conclude that log runtime for complete enumeration methods is roughly linear in k, as expected. Using the FLS data and a best subset regression approach which uses a leaps and bounds algorithm (see Section 3.2) to avoid complete enumeration of all models, he finds that the best model for the FLS data has 22 (using C_p) or 23 (using BIC) variables. These are larger model sizes than indicated by typical BMA results on these data[13].

10. Bayesian variable selection methods based on decision-theory. Often such methods avoid specifying a prior on model space and employ a utility or loss function defined on an all-encompassing model, *i.e.* a model that nests all models being considered. An early contribution is Lindley (1968), who proposes to include costs in the utility function for adding covariates, while Brown et al. (1999) extend this idea to multivariate regression. Other Bayesian model selection procedures that are based on optimising some loss or utility function can be found in *e.g.* Gelfand and Ghosh (1998), Draper and Fouskakis (2000) and Dupuis and Robert (2003). Note that decision-based approaches do need the specification of a utility function, which is arguably at least as hard to formulate as a model space prior.

11. Bayesian model averaging, discussed here in detail in Section 3.

12. Frequentist model averaging, discussed in Section 4.

In this list, methods 5-8 were specifically motivated by and introduced in economics. Note that all but the last two methods do not involve model averaging and essentially aim at uncovering a single "best" model (or a set of models for MCS). In other words, they are "model selection" methods, as opposed to the model averaging methods that we focus on here. As it is unlikely that reality (certainly in the social sciences) can be adequately captured by a simple linear model, it is often quite risky to rely on a single model for inference, forecasts and (policy) conclusions. It

[12] Mallows' C_p was developed for selecting a subset of regressors in linear regression problems. For model M_j with k_j parameters $C_p = \frac{SSE_j}{\hat{\sigma}^2} - n + 2k_j$ where SSE_j is the error sum of squares from M_j and $\hat{\sigma}^2$ the estimated error variance. $E(C_p) = k_j$ (approximately) and regressions with low C_p are favoured.

[13] For example, using the prior setup later described in (6) with fixed g, Ley and Steel (2009b) find the models with highest posterior probability to have between 5 and 10 regressors for most prior choices. Using random g, the results in Ley and Steel (2012) indicate that a typical average model size is between 10 and 20.

is much more likely[14] that an averaging method gives a better approximation to reality and it will almost certainly improve our estimate of the uncertainty associated with our conclusions. Model selection methods simply condition on the chosen model and ignore all the evidence contained in the alternative models, thus typically leading to an underestimating of the uncertainty. Comparisons of some methods (including the method by Benjamini and Hochberg (1995) aimed at controlling the false discovery rate) can be found in Deckers and Hanck (2014) in the context of cross-sectional growth regression. BMA methods can also be used for model selection, by *e.g.* simply selecting the model with the highest posterior probability. Typically, the opposite is not true as most model selection methods do not specify prior probabilities on the model space and thus can not provide posterior model probabilities.

Wang et al. (2009) claim that there are model selection methods that automatically incorporate model uncertainty by selecting variables and estimating parameters simultaneously. Such approaches are *e.g.* the SCAD penalized regression of Fan and Li (2001) and adaptive LASSO methods as in Zou (2006). These methods sometimes possess the so-called oracle property[15]. However, the oracle property is asymptotic and assumes that the "true" model is one of the models considered. So in the much more relevant context of finite samples and with true models (if they can even be formulated) outside the model space these procedures will very likely still underestimate uncertainty.[16]

Originating in machine learning, there are a number of "ensemble" algorithms like random forests, boosting or bagging (Hastie et al., 2009). As these methods typically exchange the neat, possibly structural, interpretability of a simple linear specification for the flexibility of nonlinear and nonparametric models and cannot provide probability-based uncertainty intervals, we do not consider them in this article. Nevertheless, they do often provide good predictive performance, especially in classification problems[17]. An intermediate method was proposed in Hernández et al. (2017), who combine elements of both Bayesian additive regression trees and random forests, to

[14]Strictly speaking, the choice between model averaging and model selection is related to the decision problem that we aim to solve. In most typical situations, however, the implicit loss function we specify will lead to model averaging. Examples are where we are interested in maximizing accuracy of prediction or of estimation of covariate effects.

[15]The oracle property implies that an estimating procedure identifies the "true" model asymptotically if the latter is part of the model space and has the optimal square root convergence rate. See Fan and Li (2001).

[16]For example, George (1999a) states that "BMA is well suited to yield predictive improvements over single selected models when the entire model class is misspecified. In a sense, the mixture model elaboration is an expansion of the model space to include adaptive convex combinations of models. By incorporating a richer class of models, BMA can better approximate models outside the model class."

[17]Domingos (2000) finds that BMA often fails to beat the machine learning methods in classification problems, and conjectures that this is a consequence of BMA "overfitting", in the sense that the sensitivity of the likelihood to small changes in the data carries over to the weights in (4).

offer a model-based algorithm which can deal with high-dimensional data.

3 Bayesian Model averaging

3.1 Bayesian Model Averaging: The prior and some properties

The natural Bayesian response to model uncertainty is Bayesian Model Averaging, as already explained in Section 2. Here, BMA methods are defined as those model averaging procedures for which the weights used in the averaging are based on exact or approximate posterior model probabilities and the parameters are integrated out for prediction, so there is a (sometimes implicit) prior for both models and model-specific parameters.

This paper mainly focuses on the most commonly used prior choices. Such prior structures, based on (6), have also has been shown (Bayarri et al., 2012) to have optimal properties in the sense that they satisfy several formal desirable criteria. In particular, these priors are measurement and group invariant and satisfy exact predictive matching.[18]

3.1.1 Priors on model parameters

When deciding on the priors for the model parameters, i.e. $p(\theta_j|M_j)$ in (2), it is important to realize that the prior needs to be proper on model-specific parameters. Indeed, any arbitrary constant in $p(\theta_j|M_j)$ will similarly affect the marginal likelihood $p(y|M_j)$ defined in (2). Thus, if this constant emanating from an improper prior multiplies $p(y|M_j)$ and not the marginal likelihoods for all other models, it clearly follows from (3) that posterior model probabilities are not determined. If the arbitrary constant relates to a parameter that is common to all models, it will simply cancel in the ratio (3), and for such parameters we can thus employ improper priors (Fernández et al., 2001a; Berger and Pericchi, 2001). In our normal linear model in (5), the common parameters are the intercept α and the variance σ^2, and the model-specific parameters are the β_js.

In this paper, we will primarily focus on the prior structure proposed by Fernández et al. (2001a), which is in line with the majority of the current literature[19]. Fernández et al. (2001a) start from a proper conjugate prior specification, but then adopt Jeffreys-style non-informative priors for α and σ^2. For the regression coefficients β_j, they propose a g-prior specification (Zellner,

[18] See Bayarri et al. (2012) for the precise definition of these criteria for "objective" model selection priors.

[19] A textbook treatment of this approach can be found in Chapter 11 of Koop (2003).

1986) for the covariance structure[20]. The prior density[21] is then as follows:

$$p(\alpha, \beta_j, \sigma \mid M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j | 0, \sigma^2 g (Z_j' Z_j)^{-1}), \tag{6}$$

where $f_N^q(\cdot|m, V)$ denotes the density function of a q-dimensional Normal distribution with mean m and covariance matrix V. It is worth pointing out that the dependence of the g-prior on the design matrix is not in conflict with the usual Bayesian precept that the prior should not involve the data, since the model in (5) is a model for y given Z_j, so we simply condition on the regressors throughout the analysis. The regression coefficients not appearing in M_j are exactly zero, represented by a prior point mass at zero. The amount of prior information requested from the user is limited to a single scalar $g > 0$, which can either be fixed or assigned a hyper-prior distribution. In addition, the marginal likelihood for each model (and thus the Bayes factor between each pair of models) can be calculated in closed form (Fernández et al., 2001a). In particular, the posterior distribution for the model parameters has an analytically known form as follows:

$$\begin{aligned} p(\beta_j \mid \alpha, \sigma, M_j) &= f_N^{k_j}(\beta_j \mid \delta (Z_j' Z_j)^{-1} Z_j' y, \sigma^2 \delta (Z_j' Z_j)^{-1}) && (7) \\ p(\alpha \mid \sigma, M_j) &= f_N^1(\alpha \mid \bar{y}, \sigma^2/n) && (8) \\ p(\sigma^{-2} \mid M_j) &= f_{Ga}\left(\sigma^{-2} \mid \frac{n-1}{2}, \frac{s_\delta}{2}\right), && (9) \end{aligned}$$

where $\delta = g/(1+g)$, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $s_\delta = \left[\delta y' Q_{X_j} y + (1-\delta)(y - \bar{y}\iota)'(y - \bar{y}\iota)\right]$ with $Q_W = I_n - W(W'W)^{-1}W'$ for a full column rank matrix W and $X_j = (\iota : Z_j)$ (assumed of full column rank, see footnote 10). Furthermore, $f_{Ga}(\cdot \mid a, b)$ is the density function of a Gamma distribution with mean a/b. The conditional independence between β_j and α (given σ) is a consequence of demeaning the regressors. After integrating out the model parameters as above, we can write the marginal likelihood as

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

where R_j^2 is the usual coefficient of determination for model M_j, defined through $1 - R_j^2 = y' Q_{X_j} y / [(y - \bar{y}\iota)'(y - \bar{y}\iota)]$, and the proportionality constant is the same for all models, including the null model. In addition, for each model M_j, the marginal posterior distribution of the regression coefficients β_j is a k_j-variate Student-t distribution with $n-1$ degrees of freedom, location $\delta(Z_j'Z_j)^{-1}Z_j'y$ (which is the mean if $n > 2$) and scale matrix $\delta s_\delta (Z_j'Z_j)^{-1}$ (and variance $\frac{\delta s_\delta}{n-3}(Z_j'Z_j)^{-1}$ if $n > 3$). The out-of-sample predictive distribution for each given model (which in a regression model will of course also depend on the covariate values associated with

[20] In line with most of the literature, in this paper g denotes a variance factor rather than a precision factor as in Fernández et al. (2001a).

[21] For the null model, the prior is simply $p(\alpha, \sigma) \propto \sigma^{-1}$.

the observations we want to predict) is also a Student-t distribution with $n-1$ degrees of freedom. Details can be found in equation (3.6) of Fernández et al. (2001a). Following (4), we can then conduct posterior or predictive inference by simply averaging these model-specific distributions using the posterior model weights computed (as in (3)) from (10) and the prior model distributions described in the next subsection.

There are a number of suggestions in the literature for the choice of fixed values for g, among which the most popular ones are:

- The unit information prior of Kass and Wasserman (1995) corresponds to the amount of information contained in one observation. For regular parametric families, the "amount of information" is defined through Fisher information. This gives us $g = n$, and leads to log Bayes factors that behave asymptotically like the BIC (Fernández et al., 2001a).
- The risk inflation criterion prior, proposed by Foster and George (1994), is based on the Risk inflation criterion (RIC) which leads to $g = k^2$ using a minimax perspective.
- The benchmark prior of Fernández et al. (2001a). They examine various choices of g depending on the sample size n or the model dimension k and recommend $g = \max(n, k^2)$.

When faced with a variety of possible prior choices for g, a natural Bayesian response is to formulate a hyperprior on g. This was already implicit in Zellner and Siow (1980) who use a Cauchy prior on the regression coefficients, corresponding to an inverse gamma prior on g. This idea was investigated further in Liang et al. (2008a), where hyperpriors on g are shown to alleviate certain paradoxes that appear with fixed choices for g. Sections 3.1.3 and 3.1.4 will provide more detail.

The g-prior is a relatively well-understood and convenient prior with nice properties, such as invariance under rescaling and translation of the covariates (and more generally, invariant to reparameterization under affine transformations), and automatic adaption to situations with near-collinearity between different covariates (Robert, 2007, p. 193). It can also be interpreted as the conditional posterior of the regression coefficients given a locally uniform prior and an imaginary sample of zeros with design matrix Z_j and a scaled error variance.

This idea of imaginary data is also related to the power prior approach (Ibrahim and Chen (2000) initially developed on the basis of the availability of historical data (*i.e.* data arising from previous similar studies). In addition, the mechanism of imaginary data forms the basis of the expected-posterior prior (Pérez and Berger, 2002). In Fouskakis and Ntzoufras (2016b) the power-conditional-expected-posterior prior is developed by combining the power prior and

the expected-posterior prior approaches for the regression parameters conditional on the error variance.

Som et al. (2015) introduce the block hyper-g/n prior for so-called "poly-shrinkage", which is a collection of ordinary mixtures of g-priors applied separately to groups of predictors. Their motivation is to avoid certain paradoxes, related to different asymptotic behaviour for different subsets of predictors. Min and Sun (2016) consider the situation of grouped covariates (occurring, for example, in ANOVA models where each factor has various levels) and propose separate g-priors for the associated groups of regression coefficients. This also circumvents the fact that in ANOVA models the full design matrix is often not of full rank.

A similar idea is used in Zhang et al. (2016) where a two-component extension of the g-prior is proposed, with each regressor being assigned one of two possible values for g. Their prior is proper by treating the intercept as part of the regression vector in the g-prior and by using a "vague" proper prior[22] on σ^2. They focus mostly on variable selection.

A somewhat different approach was advocated by George and McCulloch (1993, 1997), who use a prior on the regression coefficient which does not include point masses at zero. In particular, they propose a normal prior with mean zero on the entire k-dimensional vector of regression coefficients β given the model M_j which assigns a small prior variance to the coefficients of the variables that are "inactive"[23] in M_j and a larger variance to the remaining coefficients. In addition, their overall prior is proper and does not assume a common intercept.

Raftery et al. (1997b) propose yet another approach and use a proper conjugate[24] prior with a diagonal covariance structure for the regression coefficients (except for categorical predictors where a g-prior structure is used).

3.1.2 Priors over models

The prior $P(M_j)$ on model space is typically constructed by considering the probability of inclusion of each covariate. If the latter is the same for each variable, say w, and we assume inclusions are prior independent, then

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

[22]Note that this implies the necessity to choose the associated hyperparameters in a sensible manner, which is nontrivial as what is sensible depends on the scaling of the data.

[23]Formally, all variables appear in all models, but the coefficients of some variables will be shrunk to zero by the prior, indicating that their role in the model is negligible.

[24]Conjugate prior distributions combine analytically with the likelihood to give a posterior in the same class of distributions as the prior.

This implies that prior odds will favour larger models if $w > 0.5$ and the opposite if $w < 0.5$. For $w = 0.5$ all model have equal prior probability $1/k$. Defining model size as the number of included regressors in a model, a simple way to elicit w is through the prior mean model size, which is wk.[25] As the choice of w can have a substantial effect on the results, various authors (Brown et al., 1998; Clyde and George, 2004; Ley and Steel, 2009b; Scott and Berger, 2010) have suggested to put a Beta(a, b) hyperprior on w. This results in

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

which leads to much less informative priors in terms of model size. Ley and Steel (2009b) compare both approaches and suggest choosing $a = 1$ and $b = (k-m)/m$, where m is the chosen prior mean model size. This means that the user only needs to specify a value for m. The large differences between the priors in (12) and (11) can be illustrated by the prior odds they imply. Figure 2 compares the log prior odds induced by the fixed and random w priors, in the situation where $k = 67$ (corresponding to the growth dataset first used in Sala-i-Martin et al. (2004)) and using $m = 7, 33.5$ and 50. For fixed w, this corresponds to $w = 7/k, w = 1/2$ and $w = 50/67$ while for random w, we have used the specification of Ley and Steel (2009b). The figure displays the prior odds in favour of a model with $k_i = 10$ versus models with varying k_j.

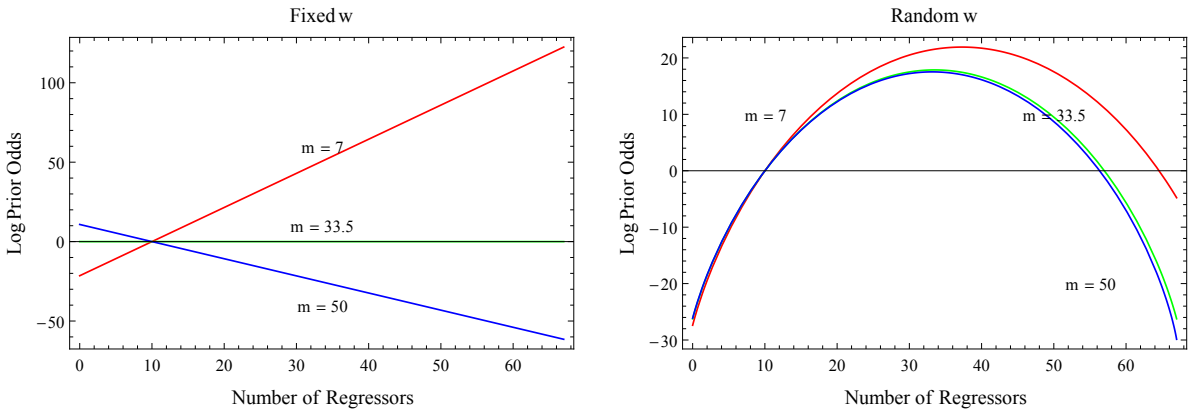

Figure 2: Log of Prior Odds: $k_i = 10$ vs varying k_j. From Ley and Steel (2009b).

Note that the random w case always leads to down-weighting 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 than of size nearer to 0 or k.[26] In contrast, the prior with fixed w does not take the number of models at each k_j into account and simply always favours larger models

[25] So, if our prior belief about mean model size is m, then we simply choose $w = m/k$.

[26] This reflects the multiplicity issue analysed more generally in Scott and Berger (2010) who propose to use (12)

when $m > k/2$ and smaller ones when $m < k/2$. Note also the much wider range of values that the log prior odds take in the case of fixed w. Thus, the choice of m is critical for the prior with fixed w, but much less so for the hierarchical prior structure, which is naturally adaptive to the data observed.

George (1999b) raises the issue of "dilution", which occurs when posterior probabilities are spread among many similar models, and suggest that prior model probabilities could have a built-in adjustment to compensate for dilution by down-weighting prior probabilities on sets of similar models. George (2010) suggests three distinct approaches for the construction of these so-called "dilution priors", based on tessellation determined neighbourhoods, collinearity adjustments, and pairwise distances between models. Dilution priors were implemented in economics by Durlauf et al. (2008) to represent priors that are uniform on theories (*i.e.* neighbourhoods of similar models) rather than on individual models, using a collinearity adjustment factor. A form of dilution prior in the context of models with interactions of covariates is the heredity prior of Chipman et al. (1997) where interaction are only allowed to be included if both main effects are included (strong heredity) or at least one of the main effects (weak heredity). In the context of examining the sources of growth in Africa, Crespo Cuaresma (2011) comments that the use of a strong heredity prior leads to different conclusions than the use of a uniform prior in the original paper by Masanjala and Papageorgiou (2008).[27] Either prior is, of course, perfectly acceptable, but it is clear that the user needs to reflect which one best captures the user's own prior ideas and intended interpretation of the results. Using the same data, Moser and Hofmarcher (2014) compare a uniform prior with a strong heredity prior and a tesselation dilution prior and find quite similar predictive performance (as measured by LPS and CRPS, explained in Section 3.1.5) but large differences in posterior inclusion probabilities (probably related to the fact that both types of dilution priors are likely to have quite different responses to multicollinearity).

Womack et al. (2015) propose viewing the model space as a partially ordered set. When the number of covariates increases, an isometry argument leads to the Poisson distribution as the unique, natural limiting prior over model dimension. This limiting prior is derived using two constructions that view an individual model as though it is a "local" null hypothesis and compares its prior probability to the probability of the alternatives that nest it. They show that this prior induces a posterior that concentrates on a finite true model asymptotically.

Another interesting recent development is the use of a loss function to assign a model prior. Equating information loss as measured by the expected minimum Kullback-Leibler divergence

with $a = b = 1$ implying a prior mean model size of $k/2$. 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.

[27] See also Papageorgiou (2011), which is a reply to the comment by Crespo Cuaresma.

between any model and its nearest model and by the "self-information loss"[28] while adding an adjustment for complexity, Villa and Lee (2016) propose the prior $P(M_j) = \exp(-ck_j)$ for some $c > 0$. This builds on the idea of Villa and Walker (2015).

3.1.3 Empirical Bayes versus Hierarchical Priors

The prior in (6) and (11) only depends on two scalar quantities, g and w. Nevertheless, these quantities can have quite a large influence on the posterior model probabilities and it is very challenging to find a single default choice for g and w that performs well in all cases, as explained in *e.g.* Fernández et al. (2001a), Berger and Pericchi (2001) and Ley and Steel (2009b). One way of reducing the impact of such prior choices on the outcome is to use hyperpriors on w and g, which fits seamlessly with the Bayesian paradigm. Hierarchical priors on w are relatively easy to deal with and were already discussed in the previous section.

Zellner and Siow (1980) used a multivariate Cauchy prior on the regression coefficients rather than the normal prior in (6). This was inspired by the argument in Jeffreys (1961) in favour of heavy-tailed priors[29]. Since a Cauchy is a scale mixture of normals, this means that implicitly the Zellner-Siow prior uses an Inverse-Gamma$(1/2, n/2)$ prior on g.

Liang et al. (2008a) introduce the hyper-g priors, which correspond to the following family of priors:

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2} \tag{13}$$

where $a > 2$ in order to have a proper distribution for $g > 0$. This includes the priors proposed in Strawderman (1971) in the context of the normal means problem. A value of $a = 4$ was suggested by Cui and George (2008) for model selection with known σ, while Liang et al. (2008a) recommend values $2 < a \leq 4$. Feldkircher and Zeugner (2009) propose to use a hyper-g prior with a value of a that leads to the same mean of the, so-called, shrinkage factor[30] $\delta = g/(1+g)$, as the unit information or the RIC prior. Ley and Steel (2012) consider the more general class of beta priors on the shrinkage factor where a Beta(b, c) prior on δ induces the following prior on g:

$$p(g) = \frac{\Gamma(b+c)}{\Gamma(b)\Gamma(c)} g^{b-1}(1+g)^{-(b+c)}. \tag{14}$$

[28]This is a loss function (also known as the log-loss function) for probability statements, which is given by the negative log of the probability.

[29]The reason for this was the limiting behaviour of the resulting Bayes factors as we consider models with better and better fit. In this case, you would want these Bayes factors, with respect to the null model, to tend to infinity. This criterion is called "information consistency" in Bayarri et al. (2012) and its absence is termed "information paradox" in Liang et al. (2008a).

[30]The name "shrinkage factor" derives from the fact that the posterior mean of the regression coefficients for a given model is the OLS estimator times this shrinkage factor, as clearly shown in (7) and the ensuing discussion.

This is an inverted beta distribution[31] (Zellner, 1971, p. 375) which clearly reduces to the hyper-g prior in (13) for $b = 1$ and $c = (a/2) - 1$. Generally, the hierarchical prior on g implies that the marginal likelihood of a given model is not analytically known, but is the integral of (10) with respect to the prior of g. Liang et al. (2008a) propose the use of a Laplace approximation for this integral, while Ley and Steel (2012) use a Gibbs sampler approach to include g in the Markov chain Monte Carlo procedure (see footnote 40). Some authors have proposed Beta shrinkage priors as in (14) that lead to analytical marginal likelihoods by making the prior depend on the model: the robust prior of Bayarri et al. (2012) truncate the prior domain to $g > [(n+1)/(k_j + 1)] - 1$ and Maruyama and George (2011) adopt the choice $b + c = (n - k_j - 1)/2$ with $c < 1/2$. However, the truncation of the robust prior is potentially problematic for cases where n is much larger than a typical model size (as is often the case in economic applications). Ley and Steel (2012) propose to use the beta shrinkage prior in (14) with mean shrinkage equal to the one corresponding to the benchmark prior of Fernández et al. (2001a) and the second parameter chosen to ensure a reasonable prior variance. They term this the benchmark beta prior and recommend using $b = c \max\{n, k^2\}$ and $c = 0.01$.

An alternative way of dealing with the problem of selecting g and w is to resort to so-called "empirical Bayes" (EB) procedures, which use the data to suggest appropriate values to choose for w and g. Of course, this amounts to using data information in selecting the prior, so is not formally in line with the Bayesian way of thinking, which prescribes a strict separation between the information in the data being analysed and that used for the prior[32]. Often, such EB methods are adopted for reasons of convenience and because they are sometimes shown to have good properties. In particular, they provide "automatic" calibration of the prior and avoid the (relatively small) computational complications that typically arise when we adopt a hyperprior on g.

Motivated by information theory, Hansen and Yu (2001) proposed a local EB method which uses a different g for each model estimated by maximizing the marginal likelihood. George and Foster (2000) develop a global EB approach, which assumes one common g and w for all models and borrows strength from all models by estimating g and w through maximizing the marginal likelihood, averaged over all models. Liang et al. (2008a) propose specific ways of estimating g in this context.

There is some evidence in the literature regarding comparisons between fully Bayes and EB procedures: Cui and George (2008) largely favour (global) EB in the context of known σ and

[31] Also known as a gamma-gamma distribution (Bernardo and Smith, 1994, p. 120).

[32] This is essentially implicit in the fact that the prior times the likelihood should define a joint distribution on the observables and the model parameters (so that e.g. the numerator in the last expression in (2) is really $p(y, \theta_j | M_j)$ and we can use the tools of probability calculus). Incidentally, it is the prior dependence on y that creates the problem, and not on Z, as the sampling model in (5) and the entire inference is implicitly conditional on Z.

$k = n$, whereas Liang et al. (2008a) find that there is little difference between EB and fully Bayes procedures (with unknown σ and $k < n$). Scott and Berger (2010) focus on EB and fully Bayesian ways of dealing with w, which, respectively, use maximum likelihood[33] and a Beta(1,1) or uniform hyperprior on w. They remark that both fully Bayesian and EB procedures exhibit clear multiplicity adjustment: as the number of noise variables increases, the posterior inclusion probabilities of variables decrease (the analysis with fixed w shows no such adjustment; see also footnote 26). However, they highlight some theoretical differences, for example the fact that EB will assign probability one to either the full model or the null model whenever one of these models has the largest marginal likelihood. They also show rather important differences in various applications, one of which uses data on GDP growth. Overall, they recommend the use of fully Bayesian procedures.

Li and Clyde (2017) compare EB and fully Bayes procedures in the more general GLM context (see Section 3.8.1), and find that local EB does badly in simulations from the null model in that it almost always selects the full model.

3.1.4 Consistency and paradoxes

One of the desiderata in Bayarri et al. (2012) for objective model selection priors is model selection consistency (introduced by Fernández et al. (2001a)), which implies that if data have been generated by M_j, then the posterior probability of M_j should converge to unity with sample size. Fernández et al. (2001a) present general conditions for the case with non-random g and show that consistency holds for *e.g.* the unit information and benchmark priors (but not for the RIC prior). When we consider hierarchical priors on g, model selection consistency is achieved by the Zellner-Siow prior in Zellner and Siow (1980) but not by local and global EB priors nor by the hyper-g prior in Liang et al. (2008a), who therefore introduce a consistent modification, the hyper-g/n prior, which corresponds to a beta distribution on $g/(n+g)$. Consistency is shown to hold for the priors of Maruyama and George (2011), Feldkircher and Zeugner (2009) (based on the unit information prior) and the benchmark beta prior of Ley and Steel (2012).

Moreno et al. (2015) consider model selection consistency when the number of potential regressors k grows with sample size. Consistency is found to depend not only on the priors for the model parameters, but also on the priors in model space. They conclude that if $k = O(n^b)$, the unit information prior, the Zellner-Siow prior and the intrinsic prior[34] lead to consistency for

[33] This is the value of w that maximizes the marginal likelihood of w summed over model space, or $\arg\max_w p(y)$ in (3), which can be referred to as type-II maximum likelihood.

[34] Intrinsic priors were introduced to justify the intrinsic Bayes factors (Berger and Pericchi, 1996). In principle, these are often based on improper reference or Jeffreys priors and the use of a so-called minimal training sample

$0 \leq b < 1/2$ under the uniform prior over model space, while consistency holds for $0 \leq b \leq 1$ if we use a Beta(1,1) hyperprior on w in (12). Wang and Maruyama (2016) investigate Bayes factor consistency associated with the prior structure in (6) for the problem of comparing nonnested models under a variety of scenarios where model dimension grows with sample size. They show that in some cases, the Bayes factor is consistent whichever the true model is, and that in others, the consistency depends on the pseudo-distance between the models. In addition, they find that the asymptotic behaviour of Bayes factors and intrinsic Bayes factors are quite similar.

Sparks et al. (2015) consider posterior consistency for parameter estimation, rather than model selection. They consider posterior consistency under the sup vector norm (weaker than the usual l_2 norm) in situations where k grows with sample size and derive necessary and sufficient conditions for consistency under the standard g-prior, the Empirical Bayes specification of George and Foster (2000) and the hyper-g and Zellner-Siow mixture priors.

Mukhopadhyay et al. (2015) show that in situations where the true model is not one of the candidate models, the use of g-priors leads to selecting a model that is in a sense closest to the true model. In addition, the loss incurred in estimating the unknown regression function under the selected model tends to that under the true model. These results have been shown under appropriate conditions on the rate of growth of g as n grows and for both the cases when the number of potential predictors remains fixed and when $k = O(n^b)$ for some $0 < b < 1$.[35] Mukhopadhyay and Samanta (2017) extend this to the situation of mixtures of g-priors and derive consistency properties for growing k under a modification of the Zellner-Siow prior, that continue to hold for more general error distributions.

Using Laplace approximations, Xiang et al. (2016) prove that in the case of hyper-g priors with growing model sizes, the Bayes factor is consistent when $k = O(n^b)$ for some $0 < b \leq 1$, even when the true model is the null model. For the case when the true model is not the null model, they show that Bayes factors are always consistent when the true model is nested within the model under consideration, and they give conditions for the non-nested case. In the specific context of analysis-of-variance (ANOVA) models, Wang (2017) shows that the Zellner-Siow prior and the beta shrinkage prior of Maruyama and George (2011) yield inconsistent Bayes factors

to convert the improper prior to a proper posterior. The latter is then used as a prior for the remaining data, so that Bayes factors can be computed. As the outcome depends on the arbitrary choice of the minimal training sample, such Bayes factors are typically "averaged" over all possible training samples. Intrinsic priors are priors that (at least asymptotically) mimic these intrinsic Bayes factors.

[35] Unlike Moreno et al. (2015), they do not explicitly find different results for different priors on the model space, which looks like an apparent contradiction. However, their results are derived under an assumption (their (A3)) bounding the ratio of prior model probabilities. Note from our Figure 2 that ratio tends to be much smaller when we use a hyperprior on w.

when k is proportional to n due to the presence of an inconsistency region around the null model. To solve the latter inconsistency, Wang (2017) propose a variation on the hyper-g/n prior, which generalizes the prior arising from a Beta distribution on $g/(\frac{n}{k}+g)$.

Finally, consistency for the power-expected-posterior approach using independent Jeffreys baseline priors is shown by Fouskakis and Ntzoufras (2016a).

A related issue is that Bayes factors can asymptotically behave in the same way as information criteria. Kass and Wasserman (1995) investigate the relationship between BIC (see Section 2.1) and Bayes factors using unit information priors for testing non-nested hypotheses and Fernández et al. (2001a) show that log Bayes factors with $g_j = n/f(k_j)$ (with $f(\cdot)$ some function which is finite for finite arguments) tend to BIC. When k is fixed, this asymptotic equivalence to BIC extends to the Zellner-Siow and Maruyama and George (2011) priors (Wang, 2017) and also the intrinsic prior (Moreno et al., 2015).

Liang et al. (2008a) remark that analyses with fixed g tend to lead to a number of paradoxical results. They mention the Bartlett (or Lindley) paradox, which is induced by the fact that very large values of g will induce support for the null model, irrespective of the data[36]. Another paradox they explore is the information paradox, where as R_i^2 tends to one, the Bayes factor in favour of M_i versus, say, the null model does not tend to ∞ but to a constant depending on g (see also footnote 29). From (16) this latter limit is $(\frac{w}{1-w})^{k_i}(1+g)^{(n-k_i-1)/2}$. Liang et al. (2008a) show that this information paradox is resolved by local or global EB methods, but also by using hyperpriors $p(g)$ that satisfy $\int(1+g)^{(n-k_i-1)/2}p(g)\mathrm{d}g = \infty$ for all $k_i \leq k$, which is the case for the Zellner-Siow prior, the hyper-g prior and the benchmark beta priors (the latter two subject to a condition, which is satisfied in most practical cases).

3.1.5 Predictive performance

Since any statistical model will typically not eliminate uncertainty and it is important to capture this uncertainty in forecasting, it is sensible to consider probabilistic forecasts, which have become quite popular in many fields. In economics, important forecasts such as the quarterly Bank of England inflation report are presented in terms of predictive distributions, and in the field of finance the area of risk management focuses on probabilistic forecasts of portfolio values. Rather than having to condition on estimated parameters, the Bayesian framework has the advantage that predictive inference can be conducted on the basis on the predictive distribution, as in (1) where all uncertainty regarding the parameters and the model is properly incorporated. This can be used

[36]This can be seen immediately by considering (16), which behaves like a constant times $(1+g)^{(k_j-k_i)/2}$ as $g \to \infty$.

to address a genuine interest in predictive questions, but also as a model-evaluation exercise. In particular, if a model estimated on a subset of the data manages to more or less accurately predict data that were not used in the estimation of the model, that intuitively suggests satisfactory performance.

In order to make this intuition a bit more precise, scoring rules provide useful summary measures for the evaluation of probabilistic forecasts. Suppose the forecaster wishes to maximize the scoring rule. If the scoring rule is proper, the forecaster has no incentive to predict any other distribution than his or her true belief for the forecast distribution. Details can be found in Gneiting and Raftery (2007).

Two important aspects of probabilistic forecasts are calibration and sharpness. Calibration refers to the compatibility between the forecasts and the observations and is a joint property of the predictions and the events that materialize. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. Proper scoring rules address both of these issues simultaneously. Popular scoring rules, used in assessing predictive performance in the context of BMA are

- The logarithmic predictive score (LPS), which is the negative of the logarithm of the predictive density evaluated at the observation. This was introduced in Good (1952) and used in the BMA context in Madigan et al. (1995), Fernández et al. (2001a,b) and Ley and Steel (2009b).

- The continuous ranked probability score (CRPS). The CRPS measures the difference between the predicted and the observed cumulative distributions as follows[37]:

$$CRPS(Q,x) = \int_{-\infty}^{\infty} \left[Q(y) - \mathbb{1}(y \leq x)\right]^2 \mathrm{d}y, \tag{15}$$

  where Q is the predictive distribution, x is the observed outcome and $\mathbb{1}(\cdot)$ is the indicator function. CRPS was found in Gneiting and Raftery (2007) to be less sensitive to outliers than LPS and was introduced in the context of growth regressions by Eicher et al. (2011).

Simple point forecasts do not allow us to take into account the uncertainty associated with the prediction, but are popular in view of their simplicity, especially in more complicated models incorporating *e.g.* dynamic aspects or endogenous regressors. Such models are often evaluated

[37] An alternative expression is given by Gneiting and Raftery (2007) as $CRPS(Q,x) = \frac{1}{2}E_Q|X-Z| - E_Q|X-x|$, where X and Z are independent copies of a random variable with distribution function Q and finite first moment. This shows that CRPS generalizes the absolute error, to which it reduces if Q is a point forecast.

in terms of the MSFE (mean squared forecast error) or the MAFE (mean absolute forecast error) calculated with respect to a point forecast.

There is a well-established literature indicating the predictive advantages of BMA. For example, Madigan and Raftery (1994) state that BMA predicts at least as well[38] as any single model in terms of LPS and Min and Zellner (1993) show that expected squared error loss of point (predictive mean) forecasts is always minimized by BMA provided the model space includes the model that generated the data. Raftery et al. (1997a) report that predictive coverage is improved by BMA with respect to prediction based on a single model. Similar results were obtained by Fernández et al. (2001a), who also use LPS as a model evaluation criterion in order to compare various choices of g in the prior (6). Fernández et al. (2001b) find, on the basis of LPS that BMA predicts substantially better than single models (such as the model with highest posterior probability) in growth data. Ley and Steel (2009b) corroborate these findings, especially with a hyperprior on w, as used in (12). Piironen and Vehtari (2017) focus on model selection methods, but state that "From the predictive point of view, best results are generally obtained by accounting for the model uncertainty and forming the full BMA solution over the candidate models, and one should not expect to do better by selection." In the context of volatility forecasting of non-ferrous metal futures, Lyócsa et al. (2017) show that averaging of forecasts substantially improves the results, especially where the averaging is conducted through BMA.

3.2 BMA in practice: Numerical methods

One advantage of the prior structure in (6) is that integration of the model parameters can be conducted analytically, and the Bayes factor between any two given models can be computed quite easily, given g. The main computational challenge is then constituted by the typically very large model space, which makes complete enumeration impossible. In other words, we simply can not try all possible models, as there are far too many of them[39].

A first possible approach is to (drastically) reduce the number of models under consideration.

[38]This optimality holds under the assumption that the data is generated by the predictive in (1) rather than a single "true" model. George (1999a) comments that "It is tempting to criticize BMA because it does not offer better average predictive performance than a correctly specified single model. However, this fact is irrelevant when model uncertainty is present because specification of the correct model with certainty is then an unavailable procedure. In most practical applications, the probability of selecting the correct model is less than 1, and a mixture model elaboration seems appropriate."

[39]In areas such as growth economics, we may have up to $k = 100$ potential covariates. This implies a model space consisting of $K = 2^{100} = 1.26 \times 10^{30}$ models. Even with fast processors, this dramatically exceeds the number of models that can be dealt with exhaustively. In other fields, the model space can even be much larger: for example, in genetics k is the number of genes and can well be of the order of tens of thousands.

One way to do this is the Occam's window algorithm, which was proposed by Madigan and Raftery (1994) for graphical models and extended to linear regression in Raftery et al. (1997b). It uses a search strategy to weed out the models that are clearly dominated by others in terms of posterior model probabilities and models that have more likely submodels nested within them. An algorithm for finding the best models is the so-called leaps and bounds method used by Raftery (1995) for BMA, based on the all-subsets regression algorithm of Furnival and Wilson (1974). The resulting set of best models can then still be reduced further through Occam's window if required. The Occam's window and the leaps and bounds algorithms are among the methods implemented in the BMA `R` package of Raftery et al. (2010) and the leaps and bounds algorithm was used in *e.g.* Masanjala and Papageorgiou (2008) and Eicher et al. (2011).

However, this tricky issue of exploring very large model spaces is now mostly dealt with through so-called Markov chain Monte Carlo (MCMC) methods[40]. In particular, a popular strategy is to run an MCMC algorithm in model space, sampling the models that are most promising: the one that is most commonly used is a random-walk Metropolis sampler usually referred to as MC^3, introduced in Madigan and York (1995) and used in *e.g.* Raftery et al. (1997a) and Fernández et al. (2001a). On the basis of the application in Masanjala and Papageorgiou (2008), Crespo Cuaresma (2011) finds that MC^3 leads to rather different results from the leaps and bounds method, which does not seem to explore the model space sufficiently well.

The original prior in George and McCulloch (1993) is not conjugate in that the prior variance of β does not involve σ^2 (unlike (6)); this means that marginal likelihoods are not available analytically, but an MCMC algorithm can easily be implemented by a Gibbs sampler on the space of the parameters and the models. This procedure is usually denoted as Stochastic Search Variable Selection (SSVS). George and McCulloch (1997) also introduce an alternative prior which is conjugate, leading to an analytical expression for the marginal likelihoods and inference can then be conducted using an MCMC sampler over only the model space (like MC^3).

[40] Suppose we have a distribution, say π, of which we do not know the properties analytically and which is difficult to simulate from directly. MCMC methods construct a Markov chain that has π as its invariant distribution and conduct inference from the generated chain. The draws in the chain are of course correlated, but ergodic theory still forms a valid basis for inference. Various algorithms can be used to generate such a Markov chain. An important one is the Metropolis-Hastings algorithm, which takes an arbitrary Markov chain and adjusts it using a simple accept-reject mechanism to ensure the stationarity of π for the resulting Markov chain. Fairly mild conditions then ensure that the values in the realized chain actually converge to draws from π. Another well-known algorithm is the Gibbs sampler, which partitions the vector of random variables which have distribution π into components and replaces each component by a draw from its conditional distribution given the current values of all other components. Various combinations of these algorithms are also popular (*e.g.* a Gibbs sampler where one or more conditionals are not easy to draw from directly and are treated through a Metropolis-Hastings algorithm). More details can be found in *e.g.* Robert and Casella (2004) and Chib (2011).

Clyde et al. (2011) remark that while the standard algorithms MC^3 and SSVS are easy to implement, they may mix poorly when covariates are highly correlated. More advanced algorithms that utilize other proposals can then be considered, such as adaptive MCMC[41] (Nott and Kohn, 2005) or evolutionary Monte Carlo (Liang and Wong, 2000). Clyde et al. (2011) propose a Bayesian adaptive sampling algorithm (BAS), that samples models without replacement from the space of models. In particular, the probability of a model being sampled is proportional to some probability mass function with known normalizing constant. Every time a new model is sampled, one needs to account for its mass by subtracting off its probability from the probability mass function to ensure that there is no duplication and then draw a new model from the renormalized distribution. The model space is represented by a binary tree structure indicating inclusion or exclusion of each variable, and marginal posterior inclusion probabilities are set at an initial estimate and then adaptively updated using the marginal likelihoods from the sampled models.

Generic numerical methods were compared in García-Donato and Martínez-Beneito (2013), who identify two different strategies:

i) MCMC methods to sample from the posterior (3) in combination with estimation based on model visit frequencies and

ii) searching methods looking for "good" models with estimation based on renormalization (*i.e.* with weights defined by the analytic expression of posterior probabilities, such as in (16)).

Despite the fact that it may, at first sight, appear that ii) should be a more efficient strategy, they show that i) is potentially more precise than ii) which could be biased by the searching procedure. Nevertheless, implementations of ii) have lead to fruitful contributions, and a lot of the most frequently used software (see Section 6) uses this method. Of course, if the algorithm simply generates a chain through model space in line with the posterior model probabilities (such as MC^3 using the prior in (6)) then both strategies can be used to conduct inference on quantities of interest, *i.e.* to compute the model probabilities to be used in (4). Indeed, Fernández et al. (2001a) suggest the use of the correlation between posterior model probabilities based on i) and ii) as an indicator of convergence of the chain. However, some other methods only lend themselves to one of the strategies above. For example, the prior of George and McCulloch (1993) does not lead to closed form expressions for the marginal likelihood, so SVSS based on this prior necessarily follows the empirical strategy i). Examples of methods that can only use strategy ii) are BAS

[41] MCMC methods often require certain parameters (of the proposal distribution) to be appropriately tuned for the algorithm to perform well. Adaptive MCMC algorithms achieve such tuning automatically. See Atchadé and Rosenthal (2005) for an introduction.

in Clyde et al. (2011), which only samples each model once, and the implementation in Raftery (1995) based on a leaps and bound algorithm which is used only to identify the top models. These methods need to use the renormalization strategy, as model visit frequencies are not an approximation to posterior model probabilities in their case.

MC^3 uses a Metropolis sampler which proposes models from a small neighbourhood of the current model, say M_j, namely all models with one covariate less or more. Whereas this works well for moderate values of k, it is not efficient in variable selection problems with large k where we expect parsimonious models to fit the data well. This is because the standard MC^3 algorithm (using a uniform distribution on the model neighbourhood) will propose to add a covariate with probability $(k - k_j)/k$, which is close to 1 if $k >> k_j$. Therefore, the algorithm will much more frequently propose to add a variable than to delete one. However, the acceptance rate of adding a new variable is equal to the acceptance rate of deleting a variable if the chain is in equilibrium. Thus, a large number of adding moves are rejected and this leads to a low between-model acceptance rate. Brown et al. (1998) extend the MC^3 proposal by adding a "swap" move where one included and one excluded covariate are selected at random and the proposed model is the one where they are swapped. They suggest to generate a candidate model by either using the MC^3 move or the swap move. Lamnisos et al. (2009) extend this further by decoupling the MC^3 move into an "add" and a "delete" move (to avoid proposing many more additions than deletions) and uniformly at random choosing whether the candidate model is generated from an "add", a "delete" or a "swap" move. In addition, they allow for less local moves by adding, deleting or swapping more than one covariate at a time. The size of the blocks of variables used for these moves is drawn from a binomial distribution. This allows for faster exploration of the model space. In Lamnisos et al. (2013) an adaptive MCMC sampler is introduced where the success probability of the binomial distribution is tuned adaptively to generate a target acceptance probability of the proposed models. They successfully manage to deal with problems like finding genetic links to colon tumours with $n = 62$ and $k = 1224$ genes in a (more challenging) probit model context (see Section 3.8.2), where their algorithm is almost 30 times more efficient[42] than MC^3 and the adaptive Gibbs sampler of Nott and Kohn (2005). Problems with even larger k can be dealt with through more sophisticated adaptive MCMC algorithms. Griffin et al. (2017) propose such algorithms which exploit the observation that in these settings the vast majority of the inclusion indicators of the variables will be virtually uncorrelated a posteriori. They are shown to lead to orders of magnitude improvements in efficiency compared to the standard Metropolis-Hastings algorithm, and are successfully applied to an extremely challenging problem with $k =$

[42]The efficiency is here standardized by CPU time. Generally, the efficiency of a Monte Carlo method is proportional to the reciprocal of the variance of the sample mean estimator normalized by the size of the generated sample.

$22,576$ possible covariates and $n = 60$ observations.

3.3 Role of the prior

It has long been understood that the effect of the prior distribution on posterior model probabilities can be much more pronounced than its effect on posterior inference given a model (Kass and Raftery, 1995; Fernández et al., 2001a). Thus, it is important to better understand the role of the prior assumptions in BMA. While Fernández et al. (2001a) examined the effects of choosing fixed values for g, a more systematic investigation of the interplay between g and w was conducted in Ley and Steel (2009b) and Eicher et al. (2011).

From combining the marginal likelihood in (10) and the model space prior in (11), we obtain the posterior odds between models, given g and w:

$$\frac{P(M_i \,|\, y, w, g)}{P(M_j \,|\, y, w, g)} = \left(\frac{w}{1-w}\right)^{k_i - k_j} (1+g)^{\frac{k_j - k_i}{2}} \left(\frac{1 + g(1 - R_i^2)}{1 + g(1 - R_j^2)}\right)^{-\frac{n-1}{2}}. \tag{16}$$

The three factors on the right-hand side of (16) correspond to, respectively, a model size (or complexity) penalty induced by the prior odds on the model space, a model size penalty resulting from the marginal likelihood (Bayes factor) and a lack-of-fit penalty from the marginal likelihood. It is clear that for fixed g and w, the complexity penalty increases with g and decreases with w (see also the discussion in Section 3.7 and in Eicher et al. (2011)). Ley and Steel (2012) consider each of the three factors separately, and define penalties as minus the logarithm of the corresponding odds factor, which ties in well with classical information criteria, which, in some cases, correspond to the limits of log posterior odds (Fernández et al., 2001a). The complexity penalty induced by the prior odds can be in favour of the smaller or the larger model, whereas the penalties induced by the Bayes factor are always in favour of the smaller and the better fitting models.

Ley and Steel (2012) find that the choice of the hyperpriors on g and w can have a large effect on the induced penalties for model complexity but does not materially affect the impact of the relative fit of the models. They also investigate how the overall complexity penalty behaves if we integrate over g and w. Figure 3 plots the logarithm of the approximate posterior odds for M_i versus M_j as a function of k_j when fixing $k_i = 10$, for different values of the prior mean model size, m, using a beta hyperprior on w as in Ley and Steel (2009b) and the benchmark beta prior on g in (14) with $c = 0.01$. We use $n = 72$ and $k = 41$ (as in the growth data of Fernández et al. (2001b)). We contrast these graphs with those for fixed values of θ and g (corresponding to the values over which the priors are centered) as derived from (16) with $R_i^2 = R_j^2$. Whereas the log posterior odds are linear in $(k_i - k_j)$ for fixed values of θ and g, they are much less extreme for

the random θ and g case, and consistently penalize models of size around $k/2$. This reflects the multiplicity penalty (see Section 3.1.2) which is implicit in the prior and analyzed in Scott and Berger (2010) in a more general context, and in Ley and Steel (2009b) in this same setting. The behaviour is qualitatively similar to that of the prior odds in Figure 2. The difference with Figure 2 is that we now consider the complexity penalty in the posterior, which also includes an (always positive) size penalty resulting from the Bayes factor. No fixed w can induce a multiplicity correction. As in Figure 2, the (relatively arbitrary) choice of m matters very little for the case with random w (and g), whereas it makes a substantial difference if we keep w (and g) fixed.

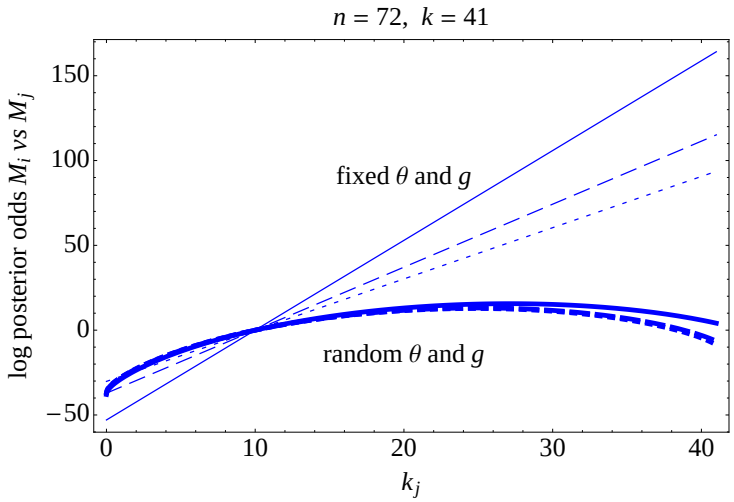

Figure 3: Posterior odds as a function of k_j when $k_i = 10$ with equal fit, using $m = 7$ (solid), $m = k/2$ (dashed), and $m = 2k/3$ (dotted). Bold lines correspond to random w and g. From Ley and Steel (2012).

Thus, marginalising out the posterior model probabilities with the hyperpriors on w and g induces a much flatter model size penalty over the entire range of model sizes. This then makes the analysis less dependent on (usually arbitrary) prior assumptions and increases the relative importance of the data contribution (the model fit) to the posterior odds.

3.4 Data Robustness

Generally, in the social sciences, the quality of the data may be problematic. An important issue is whether policy conclusions and key insights change when data are revised to eliminate errors, incorporate improved data or account for new price benchmarks. For example, the Penn World Table (PWT) income data, a dataset frequently used in cross-country empirical work in economics, have undergone periodic revisions. Ciccone and Jarociński (2010) applied the methodologies of Fernández et al. (2001b) and Sala-i-Martin et al. (2004) for investigating the determinants

of cross-country growth to data generated as in Sala-i-Martin et al. (2004) from three different versions of the PWT, versions 6.0-6.2. Both methods led to substantial variations in posterior inclusion probabilities of certain covariates between the different datasets.

It is, of course, not surprising that solving a really complicated problem (assessing the posterior distribution on a model space that contains huge quantities of models) on the basis of a very small number of observations is challenging, and if we modify the data in the absence of strong prior information, we can expect some (perhaps even dramatic) changes in our inference. Clearly, if we add prior information such changes would normally be mitigated. The perhaps most relevant question is whether we can conduct meaningful inference using BMA with the kinds of prior structures that we have discussed in this paper, such as (6).

Using the formal BMA approach, Feldkircher and Zeugner (2012) examine more in detail what causes the lack of robustness found in Ciccone and Jarociński (2010). One first conclusion is that the changes are roughly halved if the analyses with the different PWT data use the same set of countries. They also stress that the use of the fixed value of g as in the benchmark prior leads to a very large g and it is clear from (16) that this amplifies the effect of differences in the fit on posterior odds. Thus, small differences in the data can have substantial impact on the results. They propose to use a hyper-g prior which allows the model to adjust g to the data, and this dramatically reduces the instability. Interestingly, this is not a stronger prior, but a less informative one. The important thing is that fixing g at a value which is not warranted by the data quality leads to an exaggerated impact of small difference in model fit. They find that the analysis with stochastic g leads to much smaller values of g. The same behaviour was also found in Ley and Steel (2012) where three datasets were analysed: two cross-country growth datasets as in Fernández et al. (2001b) (with $n = 72$ and $k = 41$) and Sala-i-Martin et al. (2004) (with $n = 88$ and $k = 67$) and the returns-to-schooling data of Tobias and Li (2004) ($n = 1190$ and $k = 26$). In all these examples, the data favour[43] values of g in the range 15-50, which contrasts rather sharply with the fixed values of g that the benchmark prior would suggest, namely 1681, 4489 and 1190, respectively. As a consequence of the smaller g, differences between models will be less pronounced and this can be seen as a quite reasonable reaction to relatively low-quality data.

Rockey and Temple (2016) consider restricting the model space by imposing the presence of initial GDP per capita and regional dummies, *i.e.* effectively using a more informative prior on the model space. They conclude that this enhances robustness even when the analysis is extended to more recent vintages of the Penn World Table (they also consider PWT 6.3-8.0).

[43]This can be inferred from the likelihood which is marginalised with respect to all parameters but g and averaged over models; see expression (9) in Ley and Steel (2012) which is plotted in their Figure 9.

3.5 Collinearity and Jointness

One of the primary outputs of a BMA analysis is the posterior distribution of the regression coefficients, which is a mixed distribution (for each coefficient a continuous distribution with mass point at zero, reflecting exclusion of the associated regressor) of dimension k, which is almost invariably large. Thus, this is a particularly hard object to describe adequately. Summarizing this posterior distribution merely by its k marginals is obviously a gross simplification and fails to capture the truly multivariate nature of this distribution. Thus, efforts have been made to define measures that more adequately reflect the posterior distribution. Such measures should be well suited for extracting relevant pieces of information. It is important that they provide additional insight into properties of the posterior that are of particular interest, and that they are easy to interpret. Ley and Steel (2007) and Doppelhofer and Weeks (2009) propose various measures of "jointness", or the tendency of variables to appear together in a regression model. Ley and Steel (2007) formulate four desirable criteria for such measures to possess:

- Interpretability: any jointness measure should have either a formal statistical or a clear intuitive meaning in terms of jointness.
- Calibration: values of the jointness measure should be calibrated against some clearly defined scale, derived from either formal statistical or intuitive arguments.
- Extreme jointness: the situation where two variables always appear together should lead to the jointness measure reaching its value reflecting maximum jointness.
- Definition: the jointness measure should always be defined whenever at least one of the variables considered is included with positive probability.

The jointness measure proposed in Ley and Steel (2007) satisfies all of these criteria and is defined as the posterior odds ratio between those models that include a set of variables and the models that only include proper subsets. If we consider the simple case of bivariate jointness between variables i and j, and we define the events $\tilde{i}$ and $\tilde{j}$ as the exclusion of i and j, respectively, this measure is the probability of joint inclusion relative to the probability of including either regressor, but not both:

$$\mathcal{J}^{\circ}_{ij} = \frac{P(i \cap j \,|\, y)}{P(i \cap \tilde{j} \,|\, y) + P(\tilde{i} \cap j \,|\, y)}.$$

Ley and Steel (2009a) discuss how this and the other jointness measures proposed by Doppelhofer and Weeks (2009) and Strachan (2009) compare on the basis of these criteria. As $\mathcal{J}^{\circ}$ is a posterior odds ratio, its values can be immediately interpreted as evidence in favour of jointness (values above one) or disjointness (values below one, suggesting that variables are more likely to appear

on their own than jointly). Disjointness can occur, *e.g.*, when variables are highly collinear and are proxies or substitutes for each-other. In the context of two growth data sets, Ley and Steel (2007) find evidence of jointness only between important variables, which are complements in that each of them has a separate role to play in explaining growth. They find many more occurrences of disjointness, where regressors are substitutes and really should not appear together. However, these latter regressors tend to be fairly unimportant drivers of growth. Man (2017) compares different jointness measures using data from a variety of disciplines and finds that results differ substantially between the measures of Doppelhofer and Weeks (2009) on the one hand and Ley and Steel (2007) on the other hand. In contrast, results appear quite robust across different prior choices. Man (2017) suggests the use of composite indicators, which combine the information contained in the different concepts, often by simply averaging over different indicators. Given the large differences in the definitions and the properties of the jointness measures considered, I would actually expect to find considerable differences. I would recommend selecting a measure that makes sense to the user while making sure the interpretation of the results is warranted by the properties of the specific measure chosen. The use of composite indicators, however interesting it may be from the perspective of combining information, seems to me to make interpretation much harder.

Ghosh and Ghattas (2015) investigate the consequences of strong collinearity for Bayesian variable selection. They find that strong collinearity may lead to a multimodal posterior distribution over models, in which joint summaries are more appropriate than marginal summaries. They recommend a routine calculation of the joint inclusion probabilities for correlated covariates, in addition to marginal inclusion probabilities, for assessing the importance of regressors in Bayesian variable selection.

Crespo Cuaresma et al. (2016) propose a different approach to deal with patterns of inclusion such as jointness among covariates. They use a two-step approach starting from the posterior model distribution obtained from BMA methods, and then use clustering methods based on latent class analysis to unveil clusters of model profiles. Inference in the second step is based on Dirichlet process clustering methods. They also indicate that the jointness measures proposed in the literature (and mentioned earlier in this subsection) relate closely to measures used in data mining (see their footnote 1). These links are further explored in Crespo Cuaresma et al. (2017), who propose a new measure of jointness which is a regularised version of the so-called Yule's Q association coefficient, used in the machine learning literature on association rules. They use insights from the latter to extend the set of desirable criteria outlined above, and show they are satisfied by the measure they propose.

3.6 Approximations

The use of the prior structure in (6) for the linear normal model immediately leads to a closed-form marginal likelihood, but for other Bayesian models this may not be the case. In particular, the use of more complex models (such as described in Section 3.8) often do not lead to an analytic expression. One approach to addressing this problem is to use an approximation to the marginal likelihood, which is based on the ideas underlying the development of BIC (or the Schwarz criterion). In normal (or, more generally, regular[44]) models, BIC can be shown (see Schwarz, 1978 and Raftery, 1995) to provide an asymptotic approximation to the log Bayes factor. In the specific context of the normal linear model with prior (6), Fernández et al. (2001a) provide a direct link between the BIC approximation and the choice of $g = n$ (the unit information prior, which essentially leaves the asymptotics unaffected). Thus, in situations where closed-form expressions for the Bayes factors are not available (or very costly to compute), BIC has been used to approximate the actual Bayes factor. For example, some available procedures for models with endogenous regressors and models using Student-t sampling are based on BIC approximations to the marginal likelihood.

Sala-i-Martin et al. (2004) use asymptotic reasoning in the specific setting of the linear model with a g-prior to avoid specifying a prior on the model parameters and arrive at a BIC approximation in this manner. They call the resulting procedure BACE (Bayesian averaging of classical estimates). This approach was generalized to panel data by Moral-Benito (2012), who proposed Bayesian averaging of maximum likelihood estimates (BAMLE).

An alternative approximation of posterior model probabilities is through the (smoothed) AIC. Burnham and Anderson (2002) provide a Bayesian justification for AIC (with a different prior over the models than the BIC approximation) and suggest the use of AIC-based weights as posterior model probabilities. The smoothed AIC approximation is used in the context of assessing the pricing determinants of credit default swaps in Pelster and Vilsmeier (2016).

3.7 Prior robustness: illusion or not?

Previous sections have already stressed the importance of the choices of the hyperparameters and have made the point that settings for g and w are both very important for the results. However, there are examples in the literature where rather different choices of these hyperparameters led to relatively similar conclusions, which might create the impression that these choices are not that critical. For example, if we do not put hyperpriors on g and w, we note that the settings

[44]Regular models are such that the sampling distribution of the maximum likelihood estimator is asymptotically normal around the true value with covariance matrix equal to the inverse expected Fisher information matrix.

used in Fernández et al. (2001b) and in Sala-i-Martin et al. (2004) lead to rather similar results in the analysis of the growth data of Sala-i-Martin et al. (2004), which have $n = 88$ and $k = 67$. The choices made in Fernández et al. (2001b) are $w = 0.5$ and $g = k^2$, whereas the BACE analysis in Sala-i-Martin et al. (2004) is based on $w = 7/k$ (giving a prior mean model size $m = 7$) and $g = n$. As BACE attempts to avoid specifying a prior on the model parameters, the latter is not immediate, but follows from the formula used for the Bayes factors, which is essentially a close approximation to Bayes factors for the model with prior (6) using $g = n$. There is, however, an important tradeoff between values for g and w, which was visually clarified in Ley and Steel (2009b) and was also mentioned in Eicher et al. (2011). In particular, Ley and Steel (2009b) present a version of Figure 4 which shows the contours in (g, m) space of the values of fit (R_i^2) of Model M_i that would give it equal posterior probability to M_j, when $n = 88, k = 67, k_i = 8, k_j = 7$, and $R_j^2 = 0.75$.

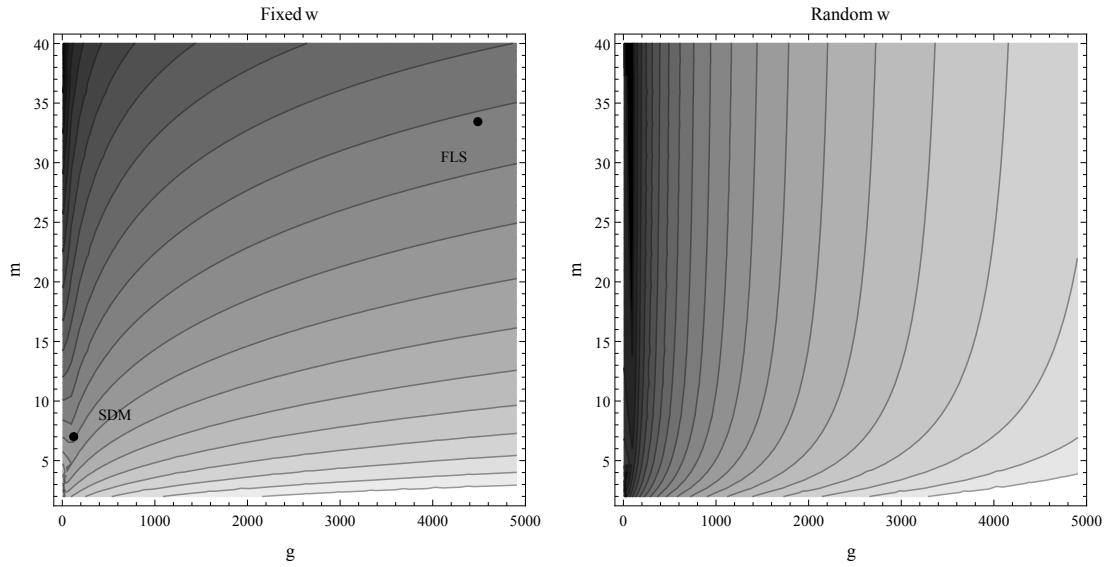

Figure 4: Equal Posterior Probability Contours for different values of R_i^2, using $n = 88, k = 67, k_i = 8, k_j = 7$, and $R_j^2 = 0.75$. The left panel is for fixed w and also indicates the choices of (g, m) in Fernández et al. (2001b) (FLS) and Sala-i-Martin et al. (2004) (SDM). The right panel corresponds to random w. Adapted from Ley and Steel (2009b).

From the left panel in Figure 4 the particular combinations of (g, w) values underlying the analyses in Fernández et al. (2001b) and in Sala-i-Martin et al. (2004) are on contours that are quite close, and thus require a very similar increase in R^2 to compensate for an extra regressor (in fact, the exact values are $R_i^2 = 0.7731$ for FLS and $R_i^2 = 0.7751$ for SDM). Remember from Section 3.3 that the model complexity penalty increases with g and decreases with w (or $m = wk$), so the effects of increasing both g and w (as in Fernández et al. (2001b) with respect

to Sala-i-Martin et al. (2004)) can cancel each-other out, as they do here.

In conclusion, it turns out that certain (often used) combinations happen to give quite similar results. However, this does not mean that results are generally robust with respect to these choices, and there is ample evidence in the literature (Ley and Steel, 2009b; Eicher et al., 2011) that these choices matter quite crucially. Also, it is important to point out that making certain prior assumptions implicit (as is done in BACE) does not mean they no longer matter. Rather, it seems to me more useful to be transparent about prior choices and to attempt to robustify the analysis by using prior structures that are less susceptible to subjectively chosen quantities. This is illustrated in the right panel of Figure 4, where the equal probability contours are drawn for the case with a Beta$(1, (k-m)/m)$ hyperprior on w. As discussed in Section 3.1.2, this prior is much less informative, which means that the actual choice of m matters much less and the trade-off between g and w has almost disappeared. A hyperprior can also be adopted for g, as in Section 3.1.3, which would further robustify the analysis (see also Section 3.3).

3.8 Other sampling models

This section describes the use of BMA in the context of other sampling models, which are sometimes fairly straightforward extensions of the normal linear regression model (for example, the Hoeting et al. (1996) model for outliers in Section 3.8.4 or the Student-t model mentioned in Section 3.8.5) and sometimes imply substantial challenges in terms of prior elicitation and numerical implementation. Many of the models below are inspired by issues arising in economics, such as dynamic models, spatial models, models for panel data and models with endogenous covariates.

3.8.1 Generalized linear models

Generalized Linear Models (GLMs) describe a more general class of models (McCullagh and Nelder, 1989) that covers the normal linear regression model but also regression models where the response variable is non-normal, such as binomial (e.g. logistic or logit regression models, probit models), Poisson, multinomial (e.g. ordered response models, proportional odds models) or gamma distributed. Sabanés Bové and Held (2011b) consider the interpretation of the g-prior in linear models as the conditional posterior of the regression coefficients given a locally uniform prior and an imaginary sample of zeros with design matrix Z_j and a scaled error variance, and extend this to the GLM context. Asymptotically, this leads to a prior which is very similar to the standard g-prior, except that it has an extra scale factor c and a weighting matrix W in the covariance structure. In many cases, $c = 1$ and $W = I$, which leads to exactly the same structure as (6). This idea was already used in the conjugate prior proposed by Chen and Ibrahim (2003),

although they only considered the case with $W = I$ and do not treat the intercept separately. For priors on g, Sabanés Bové and Held (2011b) consider a Zellner-Siow prior and a hyper-g/n prior. Both choices are shown to lead to consistent model selection in Wu et al. (2016).

The priors on the model parameters designed for GLMs in Li and Clyde (2017) employ a different type of "centering" (induced by the observed information matrix at the MLE of the coefficients), leading to a g-prior that displays local orthogonality properties at the MLE. In addition, they use a wider class of (potentially truncated) hyper-priors for g[45]. Their results rely on approximations, and, more importantly, their prior structures are data-dependent (depending on y, not just the design matrix). Interestingly, on the basis of theoretical and empirical findings in the GLM context, they recommend similar hyper-priors[46] as recommended by Ley and Steel (2012) in a linear regression setting.

The power-conditional-expected-posterior prior of Fouskakis and Ntzoufras (2016b) has also been extended to the GLM setting in Perrakis et al. (2015).

3.8.2 Probit models

A popular approach for modelling dichotomous responses uses the probit model, which is an example of a GLM. If we observe $y_1, \ldots, y_n$ taking the values either zero or one, this model assumes that the probability that $y_i = 1$ is modeled by $y_i|\eta_i \sim \text{Bernoulli}(\Phi(\eta_i))$ where Φ is the cumulative distribution function of a standard normal random variable and $\eta = (\eta_1, \eta_2, \ldots, \eta_n)'$ is a vector of linear predictors modelled as $\eta = \alpha\iota + Z_j\beta_j$, where α, β_j and Z_j are as in (5).

Typical priors have a product structure with a normal prior on β_j (for example a g-prior) and an improper uniform on α. Generally, posterior inference for the probit model can be facilitated by using the data augmentation approach of Albert and Chib (1993).

When dealing with model uncertainty, this model is often analysed through a Markov chain Monte Carlo method on the joint space of models and model parameters, since the marginal likelihood is no longer analytically available. This complicates matters with respect to the linear regression model as this space is larger than model space and the dimension of the model parameters varies with the model. Thus, reversible jump Metropolis-Hastings methods (Green, 1995) are typically used here. Details and comparison of popular algorithms can be found in Lamnisos et al. (2009).

[45] In particular, they use the class of compound confluent hypergeometric distributions, which contains most hyperpriors used in the literature as special cases.

[46] Namely, the hyper-g/n prior and the benchmark beta prior.

3.8.3 Generalized additive models

Generalized additive models are generalized linear models in which the linear predictor depends linearly on unknown smooth functions of the covariates, so these models can account for nonlinear effects; see e.g. Hastie et al. (2009). In the context of the additive model[47], Sabanés Bové and Held (2011a) consider using fractional polynomials for these smooth functions in combination with a hyper-g prior. They combine variable selection with flexible modelling of additive effects by expanding the model space to include different powers of each potential regressor. To explore this very large model space, they propose an MCMC algorithm which adapts the Occam's window strategy of Raftery et al. (1997b). Using splines for the smooth functions, Sabanés Bové et al. (2015) propose hyper-g priors based on an iterative weighted least squares approximation to the nonnormal likelihood. They conduct inference using an algorithm which is quite similar to that in Sabanés Bové and Held (2011b).

3.8.4 Outliers

The occurrence of outliers (atypical observations) is a general problem that may affect both parameter estimation and model selection, and the issue is especially relevant if the modelling assumptions are restrictive, for example by imposing normality. In the context of normal linear regression, Hoeting et al. (1996) propose a Bayesian method for simultaneous variable selection and outlier identification, using variable inflation to model outliers. They use a proper prior and recommend the use of a pre-screening procedure to generate a list of potential outliers, which are then used to define the model space to consider. Ho (2015) applies this methodology to explore the cross-country variation in the output impact of the global financial crisis in 2008-9.

Outliers are also accommodated in Doppelhofer et al. (2016). In the context of growth data, they also introduce heteroscedastic measurement error, with variance potentially differing with country and data vintage. The model also accounts for vintage fixed effects and outliers. They use data from eight vintages of the PWT (extending the data used in Sala-i-Martin et al. (2004)) to estimate the model, and conclude that 18 variables are relatively robustly associated with GDP growth over the period 1960 to 1996, even when outliers are allowed for. The quality of the data seems to improve in later vintages and varies quite a bit among the different countries. They estimate the model using JAGS, a generic MCMC software package which determines the choice of sampling strategy, but this approach is very computer-intensive[48].

[47] This is where the link function is the identity link, so we have a normally distributed response variable.

[48] They comment that a single MCMC run takes about a week to produce, even with the use of multiple computers and parallel chains.

Of course, the use of more flexible error distributions such as scale mixtures of normals (like, for example, the Student-t regression model mentioned in the next section) can be viewed as a way to make the results more robust against outliers.

3.8.5 Non-normal errors

Doppelhofer and Weeks (2011) use a Student-t model as the sampling model, instead of the normal in (5) in order to make inference more robust with respect to outliers and unmodelled heterogeneity. They consider either fixing the degrees of freedom of the Student-t or estimating it and they use the representation of a Student-t as a continuous scale mixture of normals. Throughout, they approximate posterior model probabilities by the normality-based BIC, so the posterior model probabilities remain unaffected and only the estimates of the model parameters are affected[49]. Oberdabernig et al. (2017) use a Student-t sampling model with fixed degrees of freedom in a spatial BMA framework to investigate the drivers of differences in democracy levels across countries.

Non-normality can, of course, also be accommodated by transformations of the data. Hoeting et al. (2002) combine selection of covariates with the simultaneous choice of a transformation of the dependent variable within the Box-Cox family of transformations. Charitidou et al. (2017) consider four different families of transformations along with covariate uncertainty and use model averaging based on intrinsic and fractional Bayes factors.

3.8.6 Dynamic models

In the context of simple AR(F)IMA time-series models, BMA was used in *e.g.* Koop et al. (1997).

Raftery et al. (2010) propose the idea of using state-space models in order to allow for the forecasting model to change over time while also 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. In particular, they use the following dynamic linear model, where the subscript indicates time $t = 1, \ldots, T$:

$$y_t \sim N(z_t^{(j)\prime}\theta_t^{(j)}, H^{(j)}) \tag{17}$$

$$\theta_t^{(j)} \sim N(\theta_{t-1}^{(j)}, Q_t^{(j)}), \tag{18}$$

and the superscript is the model index with models differing in the choice of covariates in the first equation. Choosing $Q_t^{(j)}$ sequences is not required as they propose to use a forgetting factor (discount factor) on the variance of the state equation (18). Using another forgetting factor, Raftery

[49] For each model they propose a simple Gibbs sampler setup after augmenting with the mixing variables.

et al. (2010) approximate the model probabilities at each point in time, which greatly simplifies the calculations. Dynamic model averaging (DMA) is where these model weights are used to average in order to conduct inference, such as predictions, and dynamic model selection (DMS) uses a single model for such inference (typically the one with the highest posterior probability) at each point in time. Koop and Korobilis (2012) apply DMA and DMS to inflation forecasting, and find that the best predictors change considerably over time and that DMA and DMS lead to improved forecasts with respect to the usual autoregressive and time-varying-parameter models. Drachal (2016) investigates the determinants of monthly spot oil prices between 1986 and 2015, using Dynamic Model Averaging (DMA) and Dynamic Model Selection (DMS). Although some interesting patterns over time were revealed, no significant evidence was found that DMA is superior in terms of forecast accuracy over, for example, a simple ARIMA model (although this seems to be based only on point forecasts, and not on predictive scores). Finally, Onorante and Raftery (2016) introduce a dynamic Occam's window to deal with larger model spaces.

van der Maas (2014) proposes a dynamic BMA framework that allows for time variation in the set of variables that is included in the model, as well as structural breaks in the intercept and conditional variance. This framework is then applied to real-time forecasting of inflation.

Other time-varying Bayesian model weight schemes are considered in Hoogerheide et al. (2010), who find that they outperform other combination forecasting schemes in terms of predictive and economic gains. They suggest forecast combinations based on a regression approach with the predictions of different models as regressors and with time-varying regression coefficients.

3.8.7 Endogeneity

If one or more of the covariates is correlated with the error term in the equation corresponding to (5), we talk of endogeneity. In particular, we consider the following extension of the model in (5):

$$y = \alpha\iota + x\gamma + Z_j\beta_j + \varepsilon \tag{19}$$

$$x = W\delta + \nu, \tag{20}$$

where x is an endogenous regressor[50] and W is a set of instruments, which are independent of ε. Finally, the error terms corresponding to observation i are identically and independently distributed as follows:

$$(\varepsilon_i, \nu_i)' \sim N(0, \Sigma), \tag{21}$$

[50]For simplicity, we focus the presentation on the case with one endogenous regressor, but this can immediately be extended.

with $\Sigma = (\sigma_{ij})$ a 2×2 covariance matrix. It is well-known that whenever $\sigma_{12} \neq 0$ this introduces a bias in the OLS estimator of γ and a standard classical approach is the use of Two-Stage Least Squares (2SLS) instead. For BMA it also leads to misleading inference on coefficients and model probabilities, even as sample size grows, as shown in Miloschewski (2016).

Tsangarides (2004) and Durlauf et al. (2008) consider the issue of endogenous regressors in a BMA context. Durlauf et al. (2008) focus on uncertainty surrounding the selection of the endogenous and exogenous variables and propose to average over 2SLS model-specific estimates for each single model. Durlauf et al. (2012) consider model averaging across just-identified models (with as many instruments as endogenous regressors). In this case, model-specific 2SLS estimates coincide with LIML estimates, which means that likelihood-based BIC weights have some formal justification.

Lenkoski et al. (2014) extend BMA to formally account for model uncertainty not only in the selection of endogenous and exogenous variables, but also in the selection of instruments. They propose a two-step procedure that first averages across the first-stage models (*i.e.* linear regressions of the endogenous variables on the instruments) and then, given the fitted endogenous regressors from the first stage, it again takes averages in the second stage. Both steps use BIC weights. Their approach was used in Eicher and Kuenzel (2016) in the context of establishing the effect of trade on growth, where feedback (and thus endogeneity) can be expected.

Koop et al. (2012) use simulated tempering to design an MCMC method that can deal with BMA in the endogeneity context in one step. It is, however, quite a complicated and computationally costly algorithm and it is nontrivial to implement.

Karl and Lenkoski (2012) propose IVBMA, which is based on the Gibbs sampler of Rossi et al. (2006) for instrumental variables models and use conditional Bayes factors to include model selection in this Gibbs algorithm. It hinges on certain restrictions (*e.g.* joint Normality of the errors is important and the prior needs to be conditionally conjugate), but the algorithm is very efficient and is implemented in an `R`-package. Jetter and Parmeter (2016) apply IVBMA to corruption determinants with a large number of endogenous regressors, and conclude that *i.a.* income levels and the extent of primary schooling emerge as important predictors.

3.8.8 Panel data and individual effects

Panel (or longitudinal) data contain information on individuals ($i = 1, \ldots, N$) over different time periods ($t = 1, \ldots, T$). Correlation between covariates and error term might arise through a time-invariant individual effect, denoted by η_i in the model

$$y_{it} = z_{it}'\beta + \eta_i + \epsilon_{it}. \tag{22}$$

Moral-Benito (2012) uses BMA in such a panel setting with strictly exogenous regressors (uncorrelated with the ϵ_{it}s but correlated with the individual effects). In this framework, the vector of regressors can also include a lagged dependent variable (y_{it-1}) which is then correlated with ϵ_{it-1}. Moral-Benito (2012) considers such a dynamic panel model within the BMA approach by combining the likelihood function discussed in Alvarez and Arellano (2003) with the unit information g-prior.

Tsangarides (2004) addresses the issues of endogenous and omitted variables by incorporating a panel data system Generalized Method of Moments (GMM) estimator. This was extended to the limited information BMA (LIBMA) approach of Mirestean and Tsangarides (2016) and Chen et al. (2017), in the context of short-T panel models with endogenous covariates using a GMM approximation to the likelihood. They then employ a BIC approximation of the limited information marginal likelihood. Moral-Benito (2016) remarks on the controversial nature of combining frequentist GMM procedures with BMA, as it is not firmly rooted in formal statistical foundations and GMM methods may require mean stationarity. Thus, Moral-Benito (2016) proposes the use of a suitable likelihood function (derived in Moral-Benito (2013)) for dynamic panel data with fixed effects and weakly exogenous[51] regressors, which is argued to be the most relevant form of endogeneity in the growth regression context. Posterior model probabilities are based on the BIC approximation of the log Bayes factors with a unit-information g-prior and a uniform prior over model space (see Section 3.6).

León-González and Montolio (2015) develop BMA methods for models for panel data with individual effects and endogenous regressors, taking into account the uncertainty regarding the choice of instruments and exogeneity restrictions. They use reversible jump MCMC methods (developed by Koop et al. (2012)) to deal with a model space that includes models that differ in the set of regressors, instruments, and exogeneity restrictions in a panel data context.

3.8.9 Spatial data

If we wish to capture spatial interactions in the data, the model for panel data in (22) can be extended to a Spatial Autoregressive (SAR) panel model as follows:

$$y_{it} = \rho \sum_{j=1}^{N} w_{ij} y_{jt} + z_{it}'\beta + \eta_i + \xi_t + \epsilon_{it}, \tag{23}$$

where $i = 1, \ldots, N$ denotes spatial location and w_{ij} is the $(i,j)^{th}$ element of the spatial weight matrix reflecting spatial proximity of the N regions, with $w_{ii} = 0$ and the matrix is normalized

[51] This implies that past shocks to the dependent variable can be correlated with current covariates, so that there is feedback from the dependent variable to the covariates

to have row-sums of unity. Finally, there are regional effects η_i and time effects $\xi_t, t = 1 \ldots, T$. BMA in this model was used in LeSage (2014), building on earlier work, such as LeSage and Parent (2007). Crespo Cuaresma et al. (2017) use SAR models to jointly model income growth and human capital accumulation and mitigate the computational requirements by using an approximation based on spatial eigenvector filtering as in Crespo Cuaresma and Feldkircher (2013). Hortas-Rico and Rios (2016) investigate the drivers of urban income inequality using Spanish municipal data. They follow the framework of LeSage and Parent (2007) to incorporate spatial effects in the BMA analysis. Piribauer and Crespo Cuaresma (2016) compare the relative performance of the BMA methods used in LeSage and Parent (2007) with two different versions of the SVSS method (see Section 3.2) for spatial autoregressive models. On simulation data the SVSS approaches tended to perform better in terms of both in-sample predictive performance and computational efficiency. Oberdabernig et al. (2017) examine democracy determinants using BMA and find that spatial spillovers are important even after controlling for a large number of geographical covariates, using a student-t version of the SAR model (with fixed degrees of freedom).

An alternative approach was proposed by Dearmon and Smith (2016), who use the nonparametric technique of Gaussian process regression to accommodate spatial patterns and develop a BMA version of this approach. They apply it to the FLS growth data augmented with spatial information.

3.8.10 Duration models

BMA methods for duration models were first examined by Volinsky et al. (1997) in the context of proportional hazard models and based on a BIC approximation. Kourtellos and Tsangarides (2015) set out to uncover the correlates of the duration of growth spells. In particular, they investigate the relationship between inequality, redistribution, and the duration of growth spells in the presence of other possible determinants. They employ BMA for Cox hazards models and extend the BMA method developed by Volinsky et al. (1997) to allow for time-dependent covariates in order to properly account for the time-varying feedback effect of the variables on the duration of growth spells. Traczynski (2017) uses a Bayesian model-averaging approach for predicting firm bankruptcies and defaults at a 12-month horizon using hazard models. The analysis is based on a Laplace approximations for the marginal likelihood, arising from the logistic likelihood and a g-prior. On model space, a collinearity-adjusted dilution prior is chosen. Exact BMA methodology was used to identify risk factors associated with dropout and delayed graduation in higher education in Vallejos and Steel (2017), who employ a discrete time competing risks survival model, dealing simultaneously with university outcomes and its associated temporal component. For

each choice of regressors, this amounts to a multinomial logistic regression model, which is a special case of a GLM. They use the prior as in Sabanés Bové and Held (2011b) in combination with the hyper-g/n prior of Liang et al. (2008b).

4 Frequentist model averaging

Frequentist methods[52] are inherently different to Bayesian methods, as they tend to focus on estimators and their properties (often, but not always, in an asymptotic setting) and do not require the specification of a prior on the parameters. Instead, parameters are treated as fixed, yet unknown, and are not assigned any probabilistic interpretation associated with prior knowledge or learning from data. Whereas Bayesian inference on parameters typically centers around the uncertainty (captured by a full posterior distribution) that remains after observing the sample in question, frequentist methods usually focus on estimators that have desirable properties in the context of repeated sampling from a given experiment.

Early examples of Frequentist Model Averaging (FMA) can be found in the forecasting literature, such as the forecast combinations of Bates and Granger (1969). This literature on forecast combinations has become quite voluminous, see *e.g.* Granger (1989) and Stock and Watson (2006) for reviews, while useful surveys of FMA can be found in Wang et al. (2009) and Burnham and Anderson (2002).

In the context of the linear regression model in (5), FMA estimators can be described as

$$\hat{\beta}_{FMA} = \sum_{j=1}^{K} \omega_j \hat{\beta}_j, \tag{24}$$

where $\hat{\beta}_j$ is an estimator based on model j and $\omega_j, j = 1 \ldots, K$ are weights in the unit simplex within $\Re^K$. The critical choice is then how to choose the weights.

Buckland et al. (1997) construct weights based on different information criteria. They propose using

$$\omega_j = \frac{\exp(-I_j/2)}{\sum_{i=1}^{K} \exp(-I_i/2)}, \tag{25}$$

where I_j is an information criterion for model j, which can be the AIC or the BIC. Burnham and Anderson (2002) recommend the use of a modified AIC criterion, which has an additional small-sample second order bias correction term. They argue that this modified AIC should be used whenever $n/k < 40$.

[52] This is the "classical" statistical methodology which still underlies most introductory textbooks in statistics and econometrics.

Hjort and Claeskens (2003) build a general large-sample likelihood framework to describe limiting distributions and risk properties of estimators post-selection as well as of model averaged estimators. Their approach also explicitly takes modeling bias into account. Besides suggesting various FMA procedures (based on e.g. AIC, the focused information criterion, FIC, of Claeskens and Hjort (2003) and empirical Bayes ideas), they provide a frequentist view of the performance of BMA schemes (in the sense of limiting distributions and large sample approximations to risks).

Hansen (2007) proposed a least squares model averaging estimator with model weights selected by minimizing the Mallows' criterion (C_p). This estimator, known as Mallows model averaging (MMA), is easily implementable for linear regression models and has certain asymptotic optimality properties, since the Mallows' criterion is asymptotically equivalent to the squared error. Therefore, the MMA estimator minimizes the squared error in large samples. Hansen (2007) shows that the weight vector chosen by MMA achieves optimality in the sense conveyed by Li (1987).

Hansen and Racine (2012) introduced another estimator within the FMA framework called jackknife model averaging (JMA) that selects appropriate weights for averaging models by minimizing a cross-validation (leave-one-out) criterion. JMA is asymptotically optimal in the sense of reaching the lowest possible squared errors over the class of linear estimators. Unlike MMA, JMA has optimality properties under heteroscedastic errors and when the candidate models are non-nested.

Liu (2015) derives the limiting distributions of least squares averaging estimators for linear regression models in a local asymptotic framework. The averaging estimators with fixed weights are shown to be asymptotically normal and a plug-in averaging estimator is proposed that minimizes the sample analog of the asymptotic mean squared error. This estimator is compared with the FIC, MMA and JMA estimators. The asymptotic distributions of averaging estimators with data-dependent weights are shown to be nonstandard and a simple procedure to construct valid confidence intervals is proposed.

Liu et al. (2016) extend MMA to linear regression models with heteroscedastic errors, and propose a model averaging method that combines generalized least squares (GLS) estimators. They derive C_p-like criteria to determine the model weights and show they are optimal in the sense of asymptotically achieving the smallest possible MSE. They also consider feasible versions using both parametric and nonparametric estimates of the error variances. Their objective is to obtain an estimator that generates a smaller MSE, which they achieve by choosing weights to minimize an estimate of the MSE. They compare their methods with those of Magnus et al. (2011), who also average feasible GLS estimators.

Most asymptotically optimal FMA methods have been developed for linear models, but

Zhang et al. (2016) specifically consider GLMs (see Section 3.8.1) and generalized linear mixed-effects models[53] and propose weights based on a plug-in estimator of the Kullback-Leibler loss plus a penalty term. They prove asymptotic optimality for fixed or growing numbers of covariates.

FMA was used for forecasting with factor-augmented regression models in Cheng and Hansen (2015). In the context of growth theory, Sala-i-Martin (1997) uses (24), but focuses on the "level of confidence"[54], using weights that are either uniform or based on the maximized likelihood.

Another model-averaging procedure that has been proposed in Magnus et al. (2010) and reviewed in Magnus and De Luca (2016) is weighted average least squares (WALS), which can be viewed as being in between BMA and FMA. The weights it implies in (24) can be given a Bayesian justification. However, it assumes no prior on model space and thus can not produce inference on posterior model probabilities. WALS is easier to compute than BMA or FMA, but quite a bit harder to explain and inherently linked to a nested linear regression setting. Magnus and De Luca (2016) provide an in-depth description of WALS and its relation to BMA and FMA. They state: "The WALS procedure surveyed in this paper is a Bayesian combination of frequentist estimators. The parameters of each model are estimated by constrained least squares, hence frequentist. However, after implementing a semiorthogonal transformation to the auxiliary regressors, the weighting scheme is developed on the basis of a Bayesian approach in order to obtain desirable theoretical properties such as admissibility and a proper treatment of ignorance. The final result is a model-average estimator that assumes an intermediate position between strict BMA and strict FMA estimators [....] Finally we emphasize (again) that WALS is a model-average procedure, not a model-selection procedure. At the end we cannot and do not want to answer the question: which model is best? This brings with it certain restrictions. For example, WALS cannot handle jointness (Ley and Steel, 2007; Doppelhofer and Weeks, 2009). The concept of jointness refers to the dependence between explanatory variables in the posterior distribution, and available measures of jointness depend on posterior inclusion probabilities of the explanatory variables, which WALS does not provide." An extension called Hierarchical WALS was proposed in Magnus and Wang (2014) to jointly deal with uncertainty in concepts and in measurements within each concept, in the spirit of dilution priors (see Section 3.1.2).

Implementation of FMA does require some way of dealing with the potentially large number of models in (24). In the context of growth applications with large model spaces, Amini and Parmeter (2012) introduce an operational version of MMA by using the same semiorthogonal

[53]These models are GLMs with so-called random effects, e.g. effects that are subject-specific in a longitudinal or panel data context.

[54]This was defined as the maximum probability mass one side of zero for a Normal distribution centred at the estimated regression coefficient with the corresponding estimated variance.

transformations as adopted in WALS.

Wagner and Hlouskova (2015) consider frequentist model averaging for principal components augmented regressions illustrated with the FLS data set on economic growth determinants. In addition, they compare and contrast their method and findings with BMA and with the WALS approach, finding some differences but also some variables that are important in all analyses. Another comparison of different methods on growth data can be found in Amini and Parmeter (2011). They consider BMA, MMA and WALS and find that results (in as far as they can be compared: for example, MMA and WALS do not provide posterior inclusion probabilities) for three growth data sets are roughly similar.

Finally, Henderson and Parmeter (2016) use FMA techniques to deal with uncertainty in a nonparametric setting and propose a nonparametric regression estimator averaged over the choices of kernel, bandwidth selection mechanism and local-polynomial order.

4.1 Density forecast combinations

As mentioned earlier, there is a large literature in forecasting which combines forecasts from different models in an equation such as (24) to provide more stable and better-performing forecasts. Of course, the choice of weights in combination forecasting is important. For example, we could consider weighting better forecasts more heavily. In addition, time-varying weights have been suggested. Stock and Watson (2004) examine a number of weighting schemes in terms of the accuracy of point forecasts and find that forecast combinations can perform well in comparison with single models, but that the best weighting schemes are often the ones that incorporate little or no data adaptivity.

There is an increasing awareness of the importance of probabilistic or density forecasts, as described in Section 3.1.5. Thus, a recent literature has emerged on density forecast combinations or weighted linear combinations (pools) of prediction models. Density forecasts combinations were discussed in Wallis (2005) and further developed by Hall and Mitchell (2007), where the combination weights are chosen to minimize the Kullback-Leibler "distance" between the predicted and true but unknown density. The latter is equivalent to optimizing LPS as defined in Section 3.1.5. The properties of such prediction pools are examined in some detail in Geweke and Amisano (2011), who show that including models that are clearly inferior to others in the pool can substantially improve prediction. Also, they illustrate that weights are not an indication of a predictive model's contribution to log score. This approach is extended by Kapetanios et al. (2015), who allow for more general specifications of the combination weights, by letting them depend on the variable to be forecast. They specifically investigate piecewise linear weight

functions and show that estimation by optimizing LPS leads to consistency and asymptotic normality[55]. They also illustrate the advantages over density forecast combinations with constant weights using simulated and real data.

5 Applications in Economics

There is a large and rapidly growing literature where model averaging techniques are used to tackle empirical problems in economics. Before the introduction of model averaging methods, model uncertainty was typically dealt with in a less formalized manner and perhaps even simply ignored in many cases. Without attempting to be exhaustive, this chapter briefly mentions some examples of model averaging in economic problems. Most of these applications relate to macroeconomic data, since the problem of model uncertainty may be more acute when dealing with these data which typically contain relatively small samples (n only a bit larger than k)[56].

5.1 Growth regressions

Traditionally, growth theory has been an area where many potential determinants have been suggested and empirical evidence has struggled to resolve the open-endedness of the theory (see footnote 9). Early attempts at finding a solution include the use of EBA (see Section 2.1) in Levine and Renelt (1992) who investigate the robustness of the results from linear regressions and find that very few regressors pass the extreme bounds test, while Sala-i-Martin (1997) employs a less severe test based on the "level of confidence" of individual regressors averaged over models (uniformly or with weights proportional to the likelihoods). These more or less intuitive but ad-hoc approaches were precursors to a more formal treatment through BMA discussed and implemented in Brock and Durlauf (2001) and Fernández et al. (2001b).

Hendry and Krolzig (2004) present an application of general-to-specific modelling (see Section 2.1) in growth theory, as an alternative to BMA. However, there is a long list of applications in this area where BMA is used, and some examples are given below.

The question of whether energy consumption is a critical driver of economic growth is investigated in Camarero et al. (2015). This relates to an important debate in economics between

[55]Formally, this is shown for known thresholds of the piecewise linear weights, and is conjectured to hold for unknown threshold parameters.

[56]However, there are lots of examples with $n >> k$ with substantial model uncertainty; for example, Ley and Steel (2012) find that the returns to schooling data of Tobias and Li (2004) where $n = 1190$ and $k = 26$ lead to MCMC chains that visit in the order of 10^5 different models if we use the recommended priors of the type (6) with random g and (12).

competing economic theories: ecological economic theory (which considers the scarcity of resources as a limiting factor for growth) and neoclassical growth theory (where it is assumed that technological progress and substitution possibilities may serve to circumvent energy scarcity problems). There are various earlier studies that concentrate on the bivariate relationship between energy consumption and economic growth, but of course the introduction of other relevant covariates is key. In order to resolve this in a formal manner, they use the BMA framework on annual US data (both aggregate and sectoral) from 1949 to 2010, with up to 32 possible covariates. Camarero et al. (2015) find that energy consumption is an important determinant of aggregate GDP growth (but their model does not investigate whether energy consumption really appears as an endogenous regressor, so that they can not assess whether there is also feedback) and also identify energy intensity, energy efficiency, the share of nuclear power and public spending as important covariates. Sectoral results support the conclusion about the importance of energy consumption, but show some variation regarding the other important determinants.

A specific focus on the effect of measures of fiscal federalism on growth was adopted in Asatryan and Feld (2015). They conclude that, after controlling for unobserved country heterogeneity, no robust effects of federalism on growth can be found.

Man (2015) investigates whether competition in the economic and political arenas is a robust determinant of aggregate growth, and whether there exists jointness among competition variables versus other growth determinants. This study also provides a comparison with EBA and with "reasonable extreme bounds analysis", which also takes the fit of the models into account. Evidence is found for the importance and positive impact on growth of financial market competition, which appears complementary to other important growth determinants. Competition in other areas does not emerge as a driver of economic growth.

Piribauer (2016) estimates growth patterns in a spatial econometric framework, building on threshold estimation approaches (Hansen, 2000) to account for structural heterogeneity in the observations. The paper uses the prior structure by George and McCulloch (1993, 1997) with SSVS (see Section 3.2).

Lanzafame (2016) derives the natural or potential growth rates of Asian economies (using a Kalman filter on a state-space model) and investigates the determinants of potential growth rates through BMA methods (while always including some of the regressors).

The influence of trade on growth is analysed in Eicher and Kuenzel (2016), who use BMA methods while taking into account the endogeneity of trade variables, through the two-stage BMA approach of Lenkoski et al. (2014). They find that sectoral export diversity serves as a crucial growth determinant for low-income countries, an effect that weakens with the level of development.

The effect of government investment versus government consumption on growth in a period of fiscal consolidation in developed economies is analysed in Jovanovic (2017). Using BMA and a dilution prior (using the determinant of the correlation matrix), it is found that public investment is likely to have a bigger impact on GDP than public consumption in the countries with high public debt. Also, and more controversially, the (investment) multiplier is likely to be higher in countries with high public debt than in countries with lower public debt. The results suggest that fiscal consolidation should be accompanied by increased public investment.

Arin and Braunfels (2017) examine the existence of the "natural resource curse" focusing on the empirical links between oil reserves and growth. They find that oil revenues have a positive effect on growth. When they include interactions and treat them simply as additional covariates, they find that the positive effect can mostly be attributed to the interaction of institutional quality and oil revenues, which would suggest that institutional quality is a necessary condition for oil revenues to have a growth-enhancing effect. However, if they use a prior that adheres to the strong heredity principle (see Section 3.1.2), they find instead that the main effect of oil rents dominates.

5.2 Inflation and Output Forecasting

In the context of time series modelling with ARIMA and ARFIMA[57] models, BMA was used for posterior inference on impulse responses for real GNP in Koop et al. (1997).

Cogley and Sargent (2005) consider Bayesian averaging of three models for inflation using dynamic model weights. Another paper that uses time-varying BMA methods for inflation forecasting is van der Maas (2014). The related strategy of dynamic model averaging, due to Raftery et al. (2010) and described in Section 3.8.6, was used in Koop and Korobilis (2012). Forecasting inflation using BMA has also been examined in Eklund and Karlsson (2007), who propose the use of so-called predictive weights in the model averaging, rather than the standard BMA based on posterior model probabilities.

Shi (2016) models and forecasts quarterly US inflation and finds that Bayesian model averaging with regime switching leads to substantial improvements in forecast performance over simple benchmark approaches (e.g. random-walk or recursive OLS forecasts) and pure BMA or Markov switching models.

Ouysse (2016) considers point and density forecasts of monthly US inflation and output growth that are generated using principal components regression (PCR) and Bayesian model

[57] ARFIMA stands for Autoregressive Fractionally Integrated Moving Average models, which are used to allow for long memory behaviour

averaging (BMA). A comparison between 24 BMA specifications and 2 PCR ones in an out-of-sample, 10-year rolling event evaluation leads to the conclusion that PCR methods perform best for predicting deviations of output and inflation from their expected paths, whereas BMA methods perform best for predicting tail events. Thus, risk-neutral policy-makers may prefer the PCR approach, while the BMA approach would be the best option for a prudential, risk-averse forecaster.

Bencivelli et al. (2017) investigate the use of BMA for forecasting GDP relative to simple bridge models[58] and factor models. They conclude that for the euro area, BMA bridge models produce smaller forecast errors than a small-scale dynamic factor model and an indirect bridge model obtained by aggregating country-specific models.

Ductor and Leiva-Leon (2016) investigate the time-varying interdependence among the economic cycles of the major world economies since the 1980's. They use a BMA panel data approach (with the model in (22) including a time trend) to find the determinants of pairwise desynchronization between the business cycles of countries. They also use WALS and find that it indicates the same main determinants as BMA.

A probit model is used for forecasting US recession periods in Aijun et al. (2017). They use a Gibbs sampler based on SSVS (but with point masses for the coefficients of the excluded regressors), and adopt a generalized double Pareto prior (which is a scale mixture of normals) for the included regression parameters along with a dilution prior based on the correlation between the covariates. Their empirical results on monthly U.S. data (from 1959:02 until 2009:02) with 108 potential covariates suggest the method performs well relative to the main competitors.

5.3 VAR and DSGE modelling

A popular econometric framework for modelling several variables is the vector autoregressive (VAR) model. BMA methodology has been applied by Garratt et al. (2003) for probability forecasting of inflation and output growth in the context of a small long-run structural vector error-correcting model of the U.K. economy. George et al. (2008) apply BMA ideas in VARs using SSVS methods with priors which do not induce exact zero restrictions on the coefficients, as in George and McCulloch (1993). Koop and Korobilis (2016) extend this to Panel VARs where the restrictions of interest involve interdependencies between and heterogeneities across cross-sectional units.

[58] Bridge models relate information published at monthly frequency to quarterly national account data, and are used for producing timely "now-casts" of economic activity.

Feldkircher and Huber (2016) use a Bayesian VAR model to explore the international spillovers of expansionary US aggregate demand and supply shocks, and of a contractionary US monetary policy shock. They use SVSS priors and find evidence for significant spillovers, mostly transmitted through financial channels and with some notable cross regional variety.

BMA methods for the more restricted dynamic stochastic general equilibrium (DSGE) models were used in Strachan and van Dijk (2013), with a particular interest in the effects of investment-specific and neutral technology shocks. Evidence from US quarterly data from 1948-2009 suggests a break in the entire model structure around 1984, after which technology shocks appear to account for all stochastic trends. Investment-specific technology shocks seem more important for business cycle volatility than neutral technology shocks.

Koop (2017) provides an intuitive and accessible overview of these types of models.

5.4 Crises and finance

Following the work of Rose and Spiegel (2011) and the earlier BMA approach of Giannone et al. (2011), Feldkircher (2014) uses BMA to identify the main macroeconomic and financial market conditions that help explain the real economic effects of the global financial crisis of 2008-9. Feldkircher et al. (2014) focus on finding leading indicators for exchange market pressures during the crisis and their BMA results indicate that inflation plays an important aggravating role, whereas international reserves act as a mitigating factor. Early warning signals are also investigated in Christofides et al. (2016) who find that the importance of such signals is specific to the particular dimension of the crisis being examined. Ho (2015) investigates the causes of the 2008-9 crisis, using BMA, BACE and the approach of Hoeting et al. (1996) (see Section 3.8.4) to deal with outliers, and finds that the three methods lead to broadly similar results. The same question about the determinants of the 2008 crisis was addressed in Chen et al. (2017), who use a hierarchical prior structure with groups of variables (grouped according to a common theory about the origins of the crisis) and individual variables within each group. They use BMA to deal with uncertainty at both levels and find that "financial policies and trade linkages are the most relevant groups with regard to the relative macroeconomic performance of different countries during the crisis. Within the selected groups, a number of pre-existing financial proxies, along with measures of trade linkages, were significantly correlated with real downturns during the crisis. Controlling for both variable uncertainty and group uncertainty, our group variable selection approach is able to identify more variables that are significantly correlated with crisis intensity than those found in past studies that select variables individually."

The drivers of financial contagion after currency crises were investigated through BMA meth-

ods in Dasgupta et al. (2011). They use a probit model for the occurrence of a currency crisis in 54 to 71 countries for four years in the 1990s and find that institutional similarity is an important predictor of financial contagion during emerging market crises. Puy (2016) investigates the global and regional dynamics in equity and bond flows, using data on portfolio investments from international mutual funds. In addition, he finds strong evidence of global contagion. To assess the determinants of contagion, he regresses the fraction of variance of equity and bond funding attributable to the world factor on a set of 14 structural variables, using both WALS and BMA.

Moral-Benito and Roehn (2016) explore the relationship between financial market regulation and current account balances. They use a dynamic panel model and combine the BMA methodology with a likelihood-based estimator that accommodates both persistence and unobserved heterogeneity.

The use of BMA in forecasting exchange rates by Wright (2008) leads to the conclusion that BMA provides slightly better out-of-sample forecasts (measured by mean squared prediction errors) than the traditional random walk benchmark. This is confirmed by Ribeiro (2017), who also argues that a bootstrap-based method, called bumping, performs even better. Iyke (2015) analyses the real exchange rate in Mauritius using BMA. Different priors are adopted, including empirical Bayes. There are attempts to control for multicollinearity in the macro determinants using three competing model priors incorporating dilution, among which the tessellation prior and the weak heredity prior (see Section 3.1.2). Adler and Grisse (2017) examine behavioral equilibrium exchange rates models, which relate a long-run cointegration relationship between real exchange rates to fundamental macroeconomic variables, in a panel regression across currencies. They use BACE to deal with model uncertainty and find that some variables are robustly linked with real exchange rates. The introduction of fixed country effects in the models greatly improves the fit to real exchange rates over time.

BMA applied to a meta-analysis is used by Zigraiova and Havranek (2016) to investigate the relationship between bank competition and financial stability. They find some evidence of publication bias[59] but encounter no clear link between bank competition and stability, even when correcting for publication bias and potential misspecifications.

Devereux and Dwyer (2016) examine the output costs associated with 150 banking crises using cross country data for years after 1970. They use BMA to identify important determinants for output changes after crises and conclude that for high-income countries the behavior of real

[59]Generally, this is the situation that the probability of a result being reported in the literature (*i.e.* of the paper being published) depends on the sign or statistical significance of the estimated effect. In this case, the authors found some evidence that some authors of primary studies tend to discard estimates inconsistent with the competition-fragility hypothesis, one of the two main hypotheses in this area.

GDP after a banking crisis is most closely associated with prior economic conditions, where above-average changes in credit tend to be associated with larger expected decreases in real GDP. For low-income economies, the existence of a stock market and deposit insurance are linked with quicker recovery of real GDP.

Pelster and Vilsmeier (2016) use Bayesian Model Averaging to assess the pricing-determinants of credit default swaps. They use an autoregressive distributed lag model with time-invariant fixed effects and approximate posterior model probabilities on the basis of smoothed AIC. They conclude that credit default swaps price dynamics can be mainly explained by factors describing firms' sensitivity to extreme market movements, in particular variables measuring tail dependence (based on so-called dynamic copula models).

Horvath et al. (2017) explore the determinants of financial development as measured by financial depth (both for banks and stock markets), the efficiency of financial intermediaries (both for banks and stock markets), financial stability and access to finance. They use BMA to analyse financial development in 80 countries using nearly 40 different explanatory variables and find that the rule of law is a major factor in influencing financial development regardless of the measure used. In addition, they conclude that the level of economic development matters and that greater wealth inequality is associated with greater stock market depth, although it does not matter for the development of the banking sector or for the efficiency of stock markets and banks.

The determinants of US monetary policy are investigated in Wölfel and Weber (2017), who conclude from a BMA analysis that over the long-run (1960-2014) the important variables in explaining the Federal Funds Rate are inflation, unemployment rates and long-term interest rates. Using samples starting in 1973 (post Bretton-Woods) and 1982 (real-time data), the fiscal deficit and monetary aggregates were also found to be relevant. Wölfel and Weber (2017) also account for parameter instability through the introduction of an unknown number of structural breaks and find strong support for models with such breaks, although they conclude that there is less evidence for structural break since the 1990s.

Watson and Deller (2017) consider the relationship between economic diversity and unemployment in the light of the economic shocks provided by the recent "Great Recession". They use a spatial BMA model allowing for spatial spillover effects on data from U.S. counties with a Herfindahl diversity index computed across 87 different sectors. They conclude that increased economic diversity within the county itself is associated with significantly reduced unemployment rates across all years of the sample (2007-2014). The economic diversity of neighbours is only strongly associated with reduced unemployment rates at the height of the Great Recession.

Ng et al. (2016) investigate the relevance of social capital in stock market development using BMA methods and conclude that trust is a robust and positive determinant of stock market depth

and liquidity.

BMA was used to identify the leading indicators of financial stress in 25 OECD countries by Vašíček et al. (2017). They find that financial stress is difficult to predict out of sample, either modelling all countries at the same time (as a panel) or individually.

5.5 Other applications

Havranek et al. (2015) use BMA in a meta-analysis of intertemporal substitution in consumption. Havranek and Sokolova (2016) investigate the mean excess sensitivity reported in studies estimating consumption Euler equations. Using BMA methods, they control for 48 variables related to the context in which researchers obtain their estimates in a sample of 2,788 estimates reported in 133 published studies. Reported mean excess sensitivity seems materially affected by demographic variables, publication bias and liquidity constraints and they conclude that the permanent income hypothesis seems a pretty good approximation of the actual behavior of the average consumer. Havranek et al. (2017) consider estimates of habit formation in consumption in 81 published studies and try and relate differences in the estimates to various characteristics of the studies. They use BMA (with MC^3) and FMA[60] and find broadly similar results using both methods. Another example of the use of BMA in meta-analysis is Philips (2016) who investigates political budget cycles and finds support for some of the context-conditional theories in that literature.

The determinants of export diversification are examined in Jetter and Ramírex Hassan (2015) who conclude that Primary school enrollment has a robust positive effect on export diversification, whereas the share of natural resources in gross domestic product lowers diversification levels. Using the IVBMA approach of Karl and Lenkoski (2012) they find that these findings are robust to accounting for endogeneity.

Kourtellos et al. (2016) use BMA methods to investigate the variation in intergenerational spatial mobility across commuter zones in the US using model priors based on the dilution idea. Their results show substantial evidence of heterogeneity, which suggests exploring nonlinearities in the spatial mobility process.

Returns to education have been examined through BMA in Tobias and Li (2004). Koop et al. (2012) use their instrumental variables BMA method in this context. Cordero et al. (2016) use

[60]Here they follow the approach suggested by Amini and Parmeter (2012), who build on Magnus et al. (2010) and use orthogonalization of the covariate space, thus reducing the number of models that need to be estimated from 2^k to k. In individual regressions they use inverse-variance weights to account for the estimated dependent variable issue.

BMA methods to assess the determinants of cognitive and non-cognitive educational outcomes in Spain.

Daude et al. (2016) investigate the drivers of productive capabilities (which are important for growth) using BACE based on bias-corrected least squares dummy variable estimates (Kiviet, 1995) in a dynamic panel context with country-specific effects.

Through a spatial BMA model, Oberdabernig et al. (2017) analyse determinants of democracy differences. Also using a model with spatial effects, Hortas-Rico and Rios (2016) examine the main drivers of urban income inequality using Spanish municipal data.

Cohen et al. (2016) investigate the social acceptance of power transmission lines using a survey that was conducted in the EU. An ordered probit model was used to model the level of acceptance and the fixed country effects of that regression were then used as dependent variables in a BMA analysis, to further explain the heterogeneity between the 27 countries covered in the survey.

In the context of production modelling through stochastic frontier models, Bayesian methods were introduced by van den Broeck et al. (1994). They deal with the uncertainty regarding the specification of the inefficiency distribution through BMA. McKenzie (2016) considers three different stochastic frontier models with varying degrees of flexibility in the dynamics of productivity change and technological growth, and uses Bayesian model averaging to conduct inference on productivity growth of railroads.

Pham (2017) investigates the impact of different globalization dimensions (both economic and non-economic) on the informal sector and shadow economy in developing countries. The methodology of León-González and Montolio (2015) is used to deal with endogenous regressors as well as country-specific fixed effects.

The effect of the abundance of resources on the efficiency of resource usage is explored in Hartwell (2016). This paper considers 130 countries over various time frames from 1970 to 2011, both resource-abundant and resource-scarce, to ascertain a link between abundance of resources and less efficient usage of those resources. Efficiency is measured by *e.g.* gas or oil consumption per unit of GDP, and 3SLS estimates are obtained for a system of equations. Model averaging is then conducted according to WALS. The paper concludes that for resource-abundant countries, the improvement of property rights will lead to a more environmentally sustainable resource usage.

Wei and Cao (2017) use dynamic model averaging (DMA) to forecast the growth rate of house prices in 30 major Chinese cities. They use the MCS test (see Section 2.1) to conclude that DMA achieves significantly higher forecasting accuracy than other models in both the recursive

and rolling forecasting modes. They find that the importance of predictors for Chinese house prices varies substantially over time and that the Google search index for house prices has recently surpassed the forecasting ability of traditional macroeconomic variables. Housing prices in Hong Kong were analysed in Magnus et al. (2011) using a GLS version of WALS.

Robust determinants of bilateral trade are investigated in Chen et al. (2017), using their LIBMA methodology (see Section 3.8.8). They find evidence of trade persistence and of important roles for the exchange rate regime, several of the traditional "core" variables of the trade gravity model and trade creation and diversion through trade agreements.

6 Software and resources

The free availability of software is generally very important for the adoption of methodology by applied users. There are a number of publicly available computational resources for conducting BMA. Early contributions are the code by Raftery et al. (1997a) (now published as an `R` package in Raftery et al. (2010)) and the `Fortran` code used by Fernández et al. (2001a). Recently, a number of `R`-packages have been created, in particular the frequently used BMS package (Feldkircher and Zeugner, 2014). Details about BMS are given in Zeugner and Feldkircher (2015). Two other well-known `R`-packages are BAS (Clyde, 2017), explained in Clyde et al. (2011), and BayesVarSel (García-Donato and Forte, 2015), described in García-Donato and Forte (2016). When endogenous regressors are suspected, the `R`-package ivbma (Lenkoski et al., 2014) implements the method of Karl and Lenkoski (2012). For situations where we wish to allow for flexible nonlinear effects of the regressors, inference for (generalized) additive models as in Sabanés Bové and Held (2011a) and Sabanés Bové et al. (2015) can be conducted by the packages glmBfp (Gravestock and Sabanés Bové, 2017) on CRAN and hypergsplines (Sabanés Bové, 2011) on R-Forge, respectively. For dynamic models, an efficient implementation of the DMA methodology of Raftery et al. (2010) is provided in the `R` package eDMA (Catania and Nonejad, 2017b), as described in Catania and Nonejad (2017a). This software uses parallel computing if shared memory multiple processors hardware is available,

In addition, code exists in other computing environments; for example LeSage (2015) describes Matlab code for BMA with spatial models. Błażejowski and Kwiatkowski (2015) present a package that implements Bayesian model averaging (including jointness measures) for gretl.[61]

Using the BMS package Amini and Parmeter (2012) successfully replicate the BMA results of Fernández et al. (2001b), Masanjala and Papageorgiou (2008) and Doppelhofer and Weeks

[61] Gretl is a free, open-source software (written in C) for econometric analysis with a graphical user interface.

(2009). Forte et al. (2017) provide a systematic review of R-packages publicly available in CRAN for Bayesian model selection and model averaging in normal linear regression models. In particular, they examine in detail the packages BAS, BayesFactor (Morey et al., 2015), BayesVarSel, BMS and mombf (Rossell et al., 2014) and highlight differences in priors that can be accommodated (within the class described in (6)), numerical implementation and posterior summaries provided. All packages lead to very similar results on a number of real data sets, and generally provide reliable inference within 10 minutes of running time on a simple PC for problems up to $p = 100$ or so covariates. They find that BAS is overall faster than the other packages considered but with a very high cost in terms of memory requirements and, overall, they recommend BAS with estimation based on model visit frequencies[62]. If memory restrictions are an issue (for moderately large p or long runs) then BayesVarSel is a good choice for small or moderate values of n, while BMS is preferable when n is large.

A number of researchers have made useful BMA resources freely available:

- Feldkircher and Zeugner: `http://bms.zeugner.eu/resources/` a dedicated resource page with lots of free software and introductory material.
- Raftery: `http://www.stat.washington.edu/raftery/Research/research.html` for his papers and `http://www.stat.washington.edu/raftery/software.html` for software and data.
- Clyde: `http://stat.duke.edu/~clyde/software` for BAS and her papers can be found at `http://www2.stat.duke.edu/~clyde/research/`.
- Steel: `http://www.warwick.ac.uk/go/msteel/steel_homepage/bma` has BMA papers that I contributed to as well as code (Fortran) and data.

A number of computational resources also exists for FMA. In particular, the R packages MuMIn (Bartoń, 2016) and AICcmodavg (Mazerolle, 2017) can handle a wide range of different models. The model confidence set approach (as described in Section 2.1) can be implemented through the R package MCS (Catania and Bernardi, 2017) as described in Bernardi and Catania (2017).

[62]The BAS package also has the option to use the sampling method (without replacement) called Bayesian Adaptive Sampling (BAS) described in Clyde et al. (2011), which is based on renormalization and leads to less accurate estimates in line with the comments in Section 3.2

7 Conclusions and recommendations

The choice between BMA versus FMA is to some extent a matter of taste and may depend on the particular focus and aims of the investigation. For this author, the theoretically optimal, finite sample nature of BMA makes it particularly attractive for use in situations of model uncertainty. Also, the availability of posterior inclusion probabilities for the regressors and the easy interpretation of model probabilities (which also allows for model selection if required) seem to be clear advantages of BMA. In addition, BMA also has important optimality properties in terms of shrinkage in high-dimensional problems. In particular, Castillo et al. (2015) prove that BMA in linear regression leads to an optimal rate of contraction of the posterior on the regression coefficients to a sparse "true" data-generating model (a model where many of the coefficients are zero), provided the prior sufficiently penalizes model complexity. Rossell and Telesca (2017) show that BMA leads to fast shrinkage for spurious coefficients (and explore so-called nonlocal priors that provide even faster shrinkage in the BMA context).

Clearly, priors matter for BMA and it is crucial to be aware of this. Looking for solutions that do not depend on prior assumptions can realistically only be achieved by hiding the implicit prior assumptions. I believe it is much preferable to be explicit about the prior assumptions and the recent research in prior sensitivity should serve to highlight which aspects of the prior are particularly critical for the results and how we can "robustify" our prior choices. A recommended way to do this is through the use of hyperpriors on hyperparameters such as w and g, given a prior structure such as the one in (6). We can then, typically, make reasonable choices for our robustified priors by eliciting simple quantities, such as prior mean model size. The resulting prior avoids being unintentionally informative and has the extra advantage of adapting to the data. For example, in cases of weak or unreliable data it will tend to favour smaller value of g, avoiding unwarranted precise distinctions between models. This may well lead to larger model sizes, but that can easily be counteracted by choosing a prior on the model space that is centered over smaller models.

Given the importance of prior assumptions, a reasonable question is whether one can assess the quality of priors or limit the array of possible choices. In principle, any coherent[63] prior which does not use the data can be seen as "valid". Nevertheless, there are legitimate questions one could (and, in my view, should) ask about the prior:

- does it adequately capture the prior beliefs of the user? Is the prior a "sensible" reflection of prior ideas, based on aspects of the model that can be interpreted? This could, for example,

[63] This means the prior is in agreement with the usual rules of probability, and prevents "Dutch book" scenarios, which would guarantee a profit in a betting setting, irrespective of the outcome.

be assessed through (transformations of) parameters or predictive quantities implied by the prior.

- does it matter for the results? If inference and decisions regarding the question of interest are not much affected over a wide range of "sensible" prior assumptions, it indicates that you need not spend a lot of time and attention to finesse these particular prior assumptions. Unfortunately, when it comes to model choice, the prior is often surprisingly important.

- what is the predictive ability (as measured by *e.g.* scoring rules)? The immediate availability of probabilistic forecasts that formally incorporate both parameter and model uncertainty provides us with a very useful tool for checking the quality of the model. If a Bayesian model predicts unobserved data well, it reflects well upon both the likelihood and the prior components of this model.

- are the desiderata of Bayarri et al. (2012) for "objective" priors satisfied? These theoretical principles, such as consistency and invariance, can be used to motivate the main prior setup in this paper.

- what are the frequentist properties? Even though frequentist arguments are, strictly speaking, not part of the underlying rationale for Bayesian procedures, these procedures often perform well in repeated sampling experiments, and BMA is not an exception[64].

Sensitivity analysis (over a range of different priors and even different sampling models) is indispensable if we want to convince our colleagues, clients and policy makers. Providing an explicit mapping from these many assumptions to the main results is a key aspect of careful applied research, and should not be neglected. There are many things that theory and prior desiderata can tell us, but there will always remain a lot that is up to the user, and then it is important to try and capture a wide array of possible reasonable assumptions underlying the analysis. In essence, this is also the key message of averaging and we should take it to heart whenever we do empirical research, certainly in non-experimental sciences such as economics.

Model uncertainty is a pervasive (and sometimes not fully recognized) problem in economic applications. BMA is a natural approach for fully taking model uncertainty into account, using simple and well-defined probabilistic arguments. We now have a good understanding about the influence of prior settings in the normal linear regression model and extensions to more complicated models have been developed. Furthermore, publicly available and well-documented

[64]However, there is no guarantee that BMA will do well in frequentist terms and, for example, there is anecdotal evidence that it can perform worse in terms of, say, mean squared error than simple least squares procedures for situations with small k.

software exists which can deal with quite large model spaces using standard computing equipment in a matter of minutes. In summary, I would strongly recommend the use of BMA with an appropriately elicited "robust" prior as a practical and easily interpretable tool for researchers dealing with economic data.

References

Adler, K. and C. Grisse (2017). Thousands of BEERs: Take your pick. *Review of International Economics*, forthcoming.

Aijun, Y., X. Ju, Y. Hongqiang, and L. Jinguan (2017). Sparse Bayesian variable selection in probit model for forecasting U.S. recessions using a large set of predictors. *Computational Economics 50*, forthcoming.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–79.

Alvarez, J. and M. Arellano (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica 71*, 1121–59.

Amini, S. and C. Parmeter (2011). Bayesian model averaging in R. *Journal of Economic and Social Measurement 36*, 253–87.

Amini, S. M. and C. F. Parmeter (2012). Comparison of model averaging techniques: Assessing growth determinants. *Journal of Applied Econometrics 27*, 870–76.

Arin, K. and E. Braunfels (2017). The resource curse revisited: A Bayesian model averaging approach. Working paper, Zayed University.

Asatryan, Z. and L. Feld (2015). Revisiting the link between growth and federalism: A Bayesian model averaging approach. *Journal of Comparative Economics 43*, 772–81.

Atchadé, Y. and J. Rosenthal (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli 11*, 815–28.

Bartoń, K. (2016). MuMIn - R package for model selection and multi-model inference. `http://mumin.r-forge.r-project.org/`.

Bates, J. and C. Granger (1969). The combination of forecasts. *Operations Research Quarterly 20*, 451–68.

Bayarri, M.-J., J. Berger, A. Forte, and G. García-Donato (2012). Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics 40*, 1550–77.

Bencivelli, L., M. Marcellino, and G. Moretti (2017). Forecasting economic activity by Bayesian bridge model averaging. *Empirical Economics*, forthcoming.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*, 289–300.

Berger, J., G. García-Donato, M. Martínez-Beneito, and V. Peña (2016). Bayesian variable selection in high dimensional problems without assumptions on prior model probabilities. arXiv 1607.02993v1.

Berger, J. and L. Pericchi (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association 91*, 109–22.

Berger, J. and L. Pericchi (2001). Objective Bayesian methods for model selection: Introduction and comparison. In P. Lahiri (Ed.), *Model Selection*, Institute of Mathematical Statistics Lecture Notes - Monograph Series 38, Beachwood, OH: IMS, pp. 135–207.

Bernardi, M. and L. Catania (2017). The model confidence set package for R. *International Journal of Computational Economics and Econometrics*, forthcoming.

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

Błażejowski, M. and J. Kwiatkowski (2015). Bayesian model averaging and jointness measures for gretl. Gretl Working Paper 2, Torun School of Banking.

Brock, W. and S. Durlauf (2001). Growth empirics and reality. *World Bank Economic Review 15*, 229–72.

Brock, W. and S. Durlauf (2015). On sturdy policy evaluation. *Journal of Legal Studies 44*, S447–73.

Brock, W., S. Durlauf, and K. West (2003). Policy evaluation in uncertain economic environments. *Brookings Papers of Economic Activity 1*, 235–322 (with discussion).

Brown, P., M. Vannucci, and T. Fearn (1998). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics 12*, 173–82.

Brown, P. J., T. Fearn, and M. Vannucci (1999). The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika 86*, 635–48.

Buckland, S., K. Burnham, and N. Augustin (1997). Model selection: an integral part of inference. *Biometrics 53*, 603–18.

Burnham, K. and D. Anderson (2002). *Model selection and multimodel inference: a practical information-theoretic approach (2nd ed.)*. New York: Springet.

Camarero, M., A. Forte, G. García-Donato, Y. Mendoza, and J. Ordoñez (2015). Variable selection in the analysis of energy consumption-growth nexus. *Energy Economics 52*, 2017–16.

Carvalho, C., N. Polson, and J. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*, 465–80.

Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015). Bayesian linear regression with sparse priors. *Annals of Statistics 43*, 1986–2018.

Catania, L. and M. Bernardi (2017). MCS: Model confidence set procedure, R package. `https://cran.r-project.org/web/packages/MCS`.

Catania, L. and N. Nonejad (2017a). Dynamic model averaging for practitioners in economics and finance: The eDMA package. *Journal of Statistical Software*, forthcoming.

Catania, L. and N. Nonejad (2017b). eDMA: Dynamic model averaging with grid search, R package. `https://cran.r-project.org/web/packages/eDMA`.

Charitidou, E., D. Fouskakis, and I. Ntzoufras (2017). Objective Bayesian transformation and variable selection using default Bayes factors. *Statistics and Computing*, forthcoming.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A 158*, 419–66 (with discussion).

Chen, H., A. Mirestean, and C. G. Tsangarides (2017). Bayesian model averaging for dynamic panels with an application to a trade gravity model. *Econometric Reviews*, forthcoming.

Chen, M. and J. Ibrahim (2003). Conjugate priors for generalized linear models. *Statistica Sinica 13*, 461–76.

Chen, R.-B., Y.-C. Chen, C.-H. Chu, and K.-J. Lee (2017). On the determinants of the 2008 financial crisis: A Bayesian approach to the selection of groups and variables. *Studies in Nonlinear Dynamics & Econometrics*, forthcoming.

Cheng, X. and B. Hansen (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics 186*, 280–93.

Chib, S. (2011). Introduction to simulation and MCMC methods. In J. Geweke, G. Koop, and H. van Dijk (Eds.), *The Oxford Handbook of Bayesian Econometrics*, Oxford: Oxford University Press, pp. 183–217.

Chipman, H., M. Hamada, and C. Wu (1997). A Bayesian variable selection approach for analyzing designed experiments with complex aliasing. *Technometrics 39*, 372–81.

Christofides, C., T. Eicher, and C. Papageorgiou (2016). Did established early warning signals predict the 2008 crises? *European Economic Review 81*, 103–114. Special issue on "Model Uncertainty in Economics".

Ciccone, A. and M. Jarociński (2010). Determinants of economic growth: Will data tell? *American Economic Journal: Macroeconomics 2*, 222–46.

Claeskens, G. and N. Hjort (2003). The focused information criterion. *Journal of the American Statistical Association 98*, 900–16.

Clyde, M. (2017). BAS: Bayesian adaptive sampling for Bayesian model averaging, R package version 1.4.3. `https://cran.r-project.org/web/packages/BAS`.

Clyde, M. and E. George (2004). Model uncertainty. *Statistical Science 19*, 81–94.

Clyde, M., J. Ghosh, and M. Littman (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics 20*, 80–101.

Cogley, T. and T. Sargent (2005). The conquest of US inflation: Learning and robustness to model uncertainty. *Review of Economic Dynamics 8*, 528–63.

Cohen, J., K. Moeltner, J. Reichl, and M. Schmidthaler (2016). An empirical analysis of local opposition to new transmission lines across the EU-27. *The Energy Journal 37*, 59–82.

Cordero, J., M. Muñiz, and C. Polo (2016). The determinants of cognitive and non-cognitive educational outcomes: empirical evidence in Spain using a Bayesian approach. *Applied Economics 48*, 3355–72.

Crespo Cuaresma, J. (2011). How different is Africa? a comment on Masanjala and Papageorgiou. *Journal of Applied Econometrics 26*, 1041–47.

Crespo Cuaresma, J., G. Doppelhofer, F. Huber, and P. Piribauer (2017). Human capital accumulation and long-term income growth projections for European regions. *Journal of Regional Science*, forthcoming.

Crespo Cuaresma, J. and M. Feldkircher (2013). Spatial filtering, model uncertainty and the speed of income convergence in Europe. *Journal of Applied Econometrics 28*, 720–741.

Crespo Cuaresma, J., B. Grün, P. Hofmarcher, S. Humer, and M. Moser (2016). Unveiling covariate inclusion structures in economic growth regressions using latent class analysis. *European Economic Review 81*, 189–202. Special issue on "Model Uncertainty in Economics".

Crespo Cuaresma, J., B. Grün, P. Hofmarcher, S. Humer, and M. Moser (2017). Let's have a joint: Measuring jointness in Bayesian model averaging. Working paper, Vienna University of Economics and Business.

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

Dasgupta, A., R. Leon-Gonzalez, and A. Shortland (2011). Regionality revisited: An examination of the direction of spread of currency crises. *Journal of International Money and Finance 30*, 831–48.

Daude, C., A. Nagengast, and J. Perea (2016). Productive capabilities: An empirical analysis of their drivers. *The Journal of International Trade & Economic Development 25*, 504–35.

Dearmon, J. and T. Smith (2016). Gaussian process regression and bayesian model averaging: An alternative approach to modeling spatial phenomena. *Geographical Analysis 48*, 82–111.

Deckers, T. and C. Hanck (2014). Variable selection in cross-section regressions: Comparisons and extensions. *Oxford Bulletin of Economics and Statistics 76*, 841–73.

Devereux, J. and G. Dwyer (2016). What determines output losses after banking crises? *Journal of International Money and Finance 69*, 69–94.

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. In *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford, CA, pp. 223–30.

Doppelhofer, G., O.-P. Moe Hansen, and M. Weeks (2016). Determinants of long-term economic growth redux: A measurement error model averaging (MEMA) approach. Working paper 19/16, Norwegian School of Economics.

Doppelhofer, G. and M. Weeks (2009). Jointness of growth determinants. *Journal of Applied Econometrics 24*, 209–44.

Doppelhofer, G. and M. Weeks (2011). Robust growth determinants. Working Paper in Economics 1117, University of Cambridge.

Drachal, K. (2016). Forecasting spot oil price in a dynamic model averaging framework - have the determinants changed over time? *Energy Economics 60*, 35–46.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B 57*, 45–97 (with discussion).

Draper, D. and D. Fouskakis (2000). A case study of stochastic optimization in health policy: Problem formulation and preliminary results. *Journal of Global Optimization 18*, 399–416.

Ductor, L. and D. Leiva-Leon (2016). Dynamics of global business cycle interdependence. *Journal of International Economics 102*, 110–27.

Dupuis, J. and C. Robert (2003). Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference 111*, 77–94.

Durlauf, S., C. Fu, and S. Navarro (2012). Assumptions matter: Model uncertainty and the deterrent effect of capital punishment. *American Economic Review: Papers and Proceedings 102*, 487–92.

Durlauf, S., A. Kourtellos, and C. Tan (2008). Are any growth theories robust? *Economic Journal 118*, 329–46.

Durlauf, S., A. Kourtellos, and C. Tan (2012). Is God in the details? a reexamination of the role of religion in economic growth. *Journal of Applied Econometrics 27*, 1059–75.

Eicher, T., C. Papageorgiou, and A. Raftery (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics 26*, 30–55.

Eicher, T. S. and D. J. Kuenzel (2016). The elusive effects of trade on growth: Export diversity and economic take-off. *Canadian Journal of Economics 49*, 264–295.

Eklund, J. and S. Karlsson (2007). Forecast combination and model averaging using predictive measures. *Econometric Reviews 26*, 329–63.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–60.

Feldkircher, M. (2014). The determinants of vulnerability to the global financial crisis 2008 to 2009: Credit growth and other sources of risk. *Journal of International Money and Finance 43*, 19–49.

Feldkircher, M., R. Horvath, and M. Rusnak (2014). Exchange market pressures during the financial crisis: A Bayesian model averaging evidence. *Journal of International Money and Finance 40*, 21–41.

Feldkircher, M. and F. Huber (2016). The international transmission of us shocks - evidence from Bayesian global vector autoregressions. *European Economic Review 81*, 167–88. Special issue on "Model Uncertainty in Economics".

Feldkircher, M. and S. Zeugner (2009). Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in Bayesian model averaging. Working Paper 09/202, IMF.

Feldkircher, M. and S. Zeugner (2012). The impact of data revisions on the robustness of growth determinants: A note on 'determinants of economic growth. will data tell'? *Journal of Applied Econometrics 27*, 686–94.

Feldkircher, M. and S. Zeugner (2014). R-package BMS: Bayesian Model Averaging in R. `http://bms.zeugner.eu`.

Fernández, C., E. Ley, and M. Steel (2001a). Benchmark priors for Bayesian model averaging. *Journal of Econometrics 100*, 381–427.

Fernández, C., E. Ley, and M. Steel (2001b). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics 16*, 563–76.

Feuerverger, A., Y. He, and S. Khatri (2012). Statistical significance of the Netflix challenge. *Statistical Science 27*, 202–31.

Forte, A., G. García-Donato, and M. Steel (2017). Methods and tools for Bayesian variable selection and model averaging in normal linear regression. Department of Statistics working paper, University of Warwick.

Foster, D. and E. George (1994). The risk inflation criterion for multiple regression. *Annals of Statistics 22*, 1947–75.

Fouskakis, D. and I. Ntzoufras (2016a). Limiting behavior of the Jeffreys power-expected-posterior Bayes factor in Gaussian linear models. *Brazilian Journal of Probability and Statistics 30*, 299–320.

Fouskakis, D. and I. Ntzoufras (2016b). Power-conditional-expected priors: Using g-priors with random imaginary data for variable selection. *Journal of Computational and Graphical Statistics 25*, 647–64.

Fragoso, T. and F. Neto (2015). Bayesian model averaging: A systematic review and conceptual classification. arXiv 1509.08864v, Universidade de São Paulo.

Furnival, G. and R. Wilson (1974). Regressions by leaps and bounds. *Technometrics 16*, 499–511.

García-Donato, G. and A. Forte (2015). BayesVarSel: Bayes factors, model choice and variable selection in linear models, R package version 1.6.1. `http://CRAN.R-project.org/package=BayesVarSel`.

García-Donato, G. and A. Forte (2016). Bayesian testing, variable selection and model averaging in linear models using R with BayesVarSel. ArXiv 1611.08118.

García-Donato, G. and M. Martínez-Beneito (2013). On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association 108*, 340–52.

Garratt, A., K. Lee, M. Pesaran, and Y. Shin (2003). Forecasting uncertainties in macroeconometric modelling: An application to the UK economy. *Journal of the American Statistical Association 98*, 829–38.

Gelfand, A. and S. Ghosh (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika 85*, 1–11.

George, E. (1999a). Comment on "Bayesian model averaging: A tutorial" by J. Hoeting, D. Madigan, A. Raftery and C. Volinksy. *Statistical Science 14*, 409–12.

George, E. (1999b). Discussion of "Bayesian model averaging and model search strategies" by M. Clyde. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 6*, Oxford: Oxford University Press, pp. 175–7.

George, E. (2010). Dilution priors: Compensating for model space redundancy. In J. Berger, T. Cai, and I. Johnstone (Eds.), *Borrowing Strength: Theory Powering Applications*, Institute of Mathematical Statistics - Collections, Vol. 6, Beachwood, OH: IMS, pp. 158–65.

George, E. and D. Foster (2000). Calibration and empirical Bayes variable selection. *Biometrika 87*, 731–47.

George, E. and R. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*, 881–89.

George, E. and R. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*, 339–73.

George, E., D. Sun, and S. Ni (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics 142*, 553–80.

Geweke, J. and G. Amisano (2011). Optimal prediction pools. *Journal of Econometrics 164*, 130–41.

Ghosh, J. and A. Ghattas (2015). Bayesian variable selection under collinearity. *The American Statistician 69*, 165–73.

Giannone, D., M. Lenza, and L. Reichlin (2011). Market freedom and the global recession. *IMF Economic Review 59*, 111–35.

Gneiting, T. and A. Raftery (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association 102*, 359–78.

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

Granger, C. (1989). Combining forecasts - twenty years later. *Journal of Forecasting 8*, 167–73.

Gravestock, I. and D. Sabanés Bové (2017). glmBfp: Bayesian fractional polynomials for GLMs, R package version 0.0-51. `https://cran.r-project.org/package=glmBfp`.

Griffin, J. and P. Brown (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis 5*, 171–88.

Griffin, J., K. Łatuszyński, and M. Steel (2017). In search of lost (mixing) time: Adaptive MCMC schemes for Bayesian variable selection with very large p. ArXiv 1708.05678, University of Warwick.

Hall, S. and J. Mitchell (2007). Combining density forecasts. *International Journal of Forecasting 23*, 1–13.

Hanck, C. (2016). I just ran two trillion regressions. *Economics Bulletin 36*, 2017–42.

Hansen, B. (2000). Sample splitting and threshold estimation. *Econometrica 68*, 575–603.

Hansen, B. (2007). Least squares model averaging. *Econometrica 75*, 1175–89.

Hansen, B. and J. Racine (2012). Jackknife model averaging. *Journal of Econometrics 157*, 38–46.

Hansen, L. and T. Sargent (2014). *Uncertainty Within Economic Models*, Volume 6 of *World Scientific Series in Economic Theory*. Singapore: World Scientific.

Hansen, M. H. and B. Yu (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association 96*, 746–74.

Hansen, P., A. Lunde, and J. Nason (2011). The model confidence set. *Econometrica 79*, 453–97.

Hartwell, C. (2016). The institutional basis of efficiency in resource-rich countries. *Economic Systems 40*, 519–38.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. New York: Springer.

Havranek, T., R. Horvath, Z. Irsova, and M. Rusnak (2015). Cross-country heterogeneity in intertemporal substitution. *Journal of International Economics 96*, 100–18.

Havranek, T., M. Rusnak, and A. Sokolova (2017). Habit formation in consumption: A meta-analysis. *European Economic Review 95*, 142–67.

Havranek, T. and A. Sokolova (2016). Do consumers really follow a rule of thumb? three thousand estimates from 130 studies say "probably not". Working Paper 8/2016, Czech National Bank.

Henderson, D. and C. Parmeter (2016). Model averaging over nonparametric estimators. In G. González-Rivera, R. Carter Hill, and T.-H. Lee (Eds.), *Essays in Honor of Aman Ullah*, Advances in Econometrics, Vol. 36, Emerald Group Publishing Limited, pp. 539–60.

Hendry, D. and H.-M. Krolzig (2005). he properties of automatic gets modelling. *Economic Journal 115*, C32–61.

Hendry, D. H. and H.-M. Krolzig (2004). We ran one regression. *Oxford Bulletin of Economics and Statistics 66*, 799–810.

Hernández, B., A. Raftery, S. Pennington, and A. Parnell (2017). Bayesian additive regression trees using Bayesian model averaging. *Statistics and Computing*, forthcoming.

Hjort, N. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*, 879–99.

Ho, T. (2015). Looking for a needle in a haystack: Revisiting the cross-country causes of the 2008-9 crisis by Bayesian model averaging. *Economica 82*, 813–40.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinksy (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*, 382–417 (with discussion).

Hoeting, J., A. Raftery, and D. Madigan (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis 22*, 251–70.

Hoeting, J., A. Raftery, and D. Madigan (2002). A method for simultaneous variable and transformation selection in linear regression. *Journal of Computational and Graphical Statistics 11*, 485–507.

Hoogerheide, L., R. Kleijn, F. Ravazzolo, H. van Dijk, and M. Verbeek (2010). Forecast accuracy and economic gains from Bayesian model averaging using time-varying weights. *Journal of Forecasting 29*, 251–69.

Hoover, K. and S. Perez (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal 2*, 167–91.

Hortas-Rico, M. and V. Rios (2016). The drivers of income inequality in cities: A spatial Bayesian Model Averaging approach. Estudios sobre la Economía Española 2016/26, FEDEA.

Horvath, R., E. Horvatova, and M. Siranova (2017). Financial development, rule of law and wealth inequality: Bayesian model averaging evidence. Discussion Paper 12/2017, Bank of Finland Institute for Economies in Transition.

Ibrahim, J. and M.-H. Chen (2000). Power prior distributions for regression models. *Statistical Science 15*, 46–60.

Iyke, B. (2015). Macro determinants of the real exchange rate in a small open small island economy: Evidence from mauritius via bma. MPRA Paper 68968.

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

Jetter, M. and C. Parmeter (2016). Uncovering the determinants of corruption. Department of Economics Working Paper Series 2016-02, University of Miami.

Jetter, M. and A. Ramírex Hassan (2015). Want export diversification? educate the kids first. *Economic Inquiry 53*, 1765–82.

Jovanovic, B. (2017). Growth forecast errors and government investment and consumption multipliers. *International Review of Applied Economics 31*, 83–107.

Kapetanios, G., J. Mitchell, S. Price, and N. Fawcett (2015). Generalised density forecast combinations. *Journal of Econometrics 188*, 150–65.

Karl, A. and A. Lenkoski (2012). Instrumental variables Bayesian model averaging via conditional Bayes factors. Technical Report arXiv: 1202.5846v3, Heidelberg University.

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

Kass, R. 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*, 928–34.

Kiviet, J. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics 68*, 53–78.

Koop, G. (2003). *Bayesian Econometrics*. Chichester: Wiley.

Koop, G. (2017). Bayesian methods for empirical macroeconomics with big data. *Review of Economic Analysis 9*, 33–56.

Koop, G. and D. Korobilis (2012). Forecasting inflation using dynamic model averaging. *International Economic Review 53*, 867–86.

Koop, G. and D. Korobilis (2016). Model uncertainty in panel vector autoregressive models. *European Economic Review 81*, 115–131. Special issue on "Model Uncertainty in Economics".

Koop, G., R. Leon-Gonzalez, and R. Strachan (2012). Bayesian model averaging in the instrumental variable regression model. *Journal of Econometrics 171*, 237–50.

Koop, G., E. Ley, J. Osiewalski, and M. Steel (1997). Bayesian analysis of long memory and persistence using ARFIMA models. *Journal of Econometrics 76*, 149–69.

Kourtellos, A., C. Marr, and C. Tan (2016). Robust determinants of intergenerational mobility in the land of opportunity. *European Economic Review 81*, 132–47. Special issue on "Model Uncertainty in Economics".

Kourtellos, A. and C. Tsangarides (2015). Robust correlates of growth spells: Do inequality and redistribution matter? Working Paper 15-20, The Rimini Centre for Economic Analysis.

Lamnisos, D., J. E. Griffin, and M. F. J. Steel (2009). Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics 18*, 592–612.

Lamnisos, D., J. E. Griffin, and M. F. J. Steel (2013). Adaptive Monte Carlo for Bayesian Variable Selection in Regression Models. *Journal of Computational and Graphical Statistics 22*, 729–48.

Lanzafame, M. (2016). Potential growth in Asia and its determinants: An empirical investigation. *Asian Development Review 33*, 1–27.

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

Leamer, E. (1983). Let's take the con out of econometrics. *American Economic Review 73*, 31–43.

Leamer, E. (1985). Sensitivity analyses would help. *American Economic Review 75*, 308–13.

Leamer, E. (2016a). S-values and Bayesian weighted all-subsets regressions. *European Economic Review 81*, 15–31. Special issue on "Model Uncertainty in Economics".

Leamer, E. (2016b). S-values: Conventional context-minimal measures of the sturdiness of regression coefficients. *Journal of Econometrics 193*, 147–61.

Lenkoski, A., T. Eicher, and A. Raftery (2014). Two-stage Bayesian model averaging in endogenous variable models. *Econometric Reviews 33*, 122–51.

Lenkoski, A., A. Karl, and A. Neudecker (2014). ivbma: Bayesian instrumental variable estimation and model determination via conditional bayes factors, R package. `https://cran.r-project.org/web/packages/ivbma`.

León-González, R. and D. Montolio (2015). Endogeneity and panel data in growth regressions: A Bayesian model averaging approach. *Journal of Macroeconomics 46*, 23–39.

LeSage, J. (2014). Spatial econometric panel data model specification: a bayesian approach. *Spatial Statistics 9*, 122–45.

LeSage, J. (2015). Software for bayesian cross section and panel spatial model comparison. *Journal of Geographical Systems 17*, 297–310.

LeSage, J. P. and O. Parent (2007). Bayesian model averaging for spatial econometric models. *Geographical Analysis 39*, 241–67.

Levine, R. and D. Renelt (1992). A sensitivity analysis of cross-country growth regressions. *American Economic Review 82*, 942–63.

Ley, E. and M. Steel (2007). Jointness in Bayesian variable selection with applications to growth regression. *Journal of Macroeconomics 29*, 476–93.

Ley, E. and M. Steel (2009a). Comments on 'Jointness of growth determinants'. *Journal of Applied Econometrics 24*, 248–51.

Ley, E. and M. Steel (2009b). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics 24*, 651–74.

Ley, E. and M. Steel (2012). Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics 171*, 251–66.

Li, K.-C. (1987). Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics 15*, 958–75.

Li, Y. and M. Clyde (2017). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, forthcoming.

Liang, F., R. Paulo, G. Molina, M. Clyde, and J. Berger (2008a). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association 103*, 410–423.

Liang, F., R. Paulo, G. Molina, M. Clyde, and J. Berger (2008b). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association 103*, 410–23.

Liang, F. and W. Wong (2000). Evolutionary Monte Carlo: Applications to c_p model sampling and change point problem. *Statistica Sinica 10*, 317–42.

Lindley, D. (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society Ser. B, 30*, 31–66.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics 186*, 142–59.

Liu, Q., R. Okui, and A. Yoshimura (2016). Generalized least squares model averaging. *Econometric Reviews 35*, 1692–752.

Lyócsa, v., P. Molnár, and N. Todorova (2017). Volatility forecasting of non-ferrous metal futures: Covariances, covariates or combinations? *Journal of International Financial Markets, Institutions & Money*, fortcoming.

Madigan, D., J. Gavrin, and A. Raftery (1995). Eliciting prior information to enhance the predictive perfromance of Bayesian graphical models. *Communications in Statistics, Theory and Methods 24*, 2271–92.

Madigan, D. and A. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association 89*, 1535–46.

Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *International Statistical Review 63*, 215–32.

Magnus, J. and G. De Luca (2016). Weighted-average least squares (WALS): A survey. *Journal of Economic Surveys 30*, 117–48.

Magnus, J. and W. Wang (2014). Concept-based bayesian model averaging and growth empirics. *Oxford Bulletin Of Economics And Statistics 76*, 874–97.

Magnus, J. R., O. Powell, and P. Prüfer (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics 154*, 139–53.

Magnus, J. R., A. T. Wan, and X. Zhang (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics and Data Analysis 55*, 1331–41.

Man, G. (2015). Competition and the growth of nations: International evidence from Bayesian model averaging. *Economic Modelling 51*, 491–501.

Man, G. (2017). Critical appraisal of jointness concepts in Bayesian model averaging: Evidence from life sciences, sociology, and other scientific fields. *Journal of Applied Statistics*, forthcoming.

Marinacci, M. (2015). Model uncertainty. *Journal of the European Economic Association 13*, 1022–1100.

Maruyama, Y. and E. George (2011). Fully Bayes factors with a generalized g-prior. *Annals of Statistics 39*, 2740–2765.

Masanjala, W. and C. Papageorgiou (2008). Rough and lonely road to prosperity: a reexamination of the sources of growth in Africa using Bayesian model averaging. *Journal of Applied Econometrics 23*, 671–82.

Mazerolle, M. (2017). AICcmodavg - model selection and multimodel inference based on (q)aic(c), R package. `https://cran.r-project.org/web/packages/AICcmodavg`.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. Chapman and Hall.

McKenzie, T. (2016). Technological change and productivity in the rail industry: A Bayesian approach. Technical report, University of Oregon.

Miloschewski, A. (2016). Model uncertainty and the endogeneity problem. 9-month PhD progress report, University of Warwick.

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.

Min, X. and D. Sun (2016). Bayesian model selection for a linear model with grouped covariates. *Annals of the Institute of Statistical Mathematics 68*, 877–903.

Mirestean, A. and C. G. Tsangarides (2016). Growth determinants revisited using limited-information Bayesian model averaging. *Journal of Applied Econometrics 31*, 106–32.

Moral-Benito, E. (2012). Determinants of economic growth: a Bayesian panel data approach. *Review of Economics and Statistics 94*, 566–79.

Moral-Benito, E. (2013). Likelihood-based estimation of dynamic panels with predetermined regressors. *Journal of Business and Economic Statistics 31*, 451–72.

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

Moral-Benito, E. (2016). Growth empirics in panel data under model uncertainty and weak exogeneity. *Journal of Applied Econometrics 31*, 584–602.

Moral-Benito, E. and O. Roehn (2016). The impact of financial regulation on current account balances. *European Economic Review 81*, 148–66. Special issue on "Model Uncertainty in Economics".

Moreno, E., J. Girón, and G. Casella (2015). Posterior model consistency in variable selection as the model dimension grows. *Statistical Science 30*, 228–41.

Morey, R. D., J. N. Rouder, and T. Jamil (2015). BayesFactor: Computation of Bayes factors for common designs, R package version 0.9.11-1. `http://CRAN.R-project.org/package=BayesFactor`.

Moser, M. and P. Hofmarcher (2014). Model priors revisited: Interaction terms in BMA growth applications. *Journal of Applied Econometrics 29*, 344–47.

Mukhopadhyay, M. and T. Samanta (2017). A mixture of g-priors for variable selection when the number of regressors grows with the sample size. *Test 26*, 377–404.

Mukhopadhyay, M., T. Samanta, and A. Chakrabarti (2015). On consistency and optimality of Bayesian variable selection based on g-prior in normal linear regression models. *Annals of the Institute of Statistical Mathematics 67*, 963–997.

Ng, A., M. Ibrahim, and A. Mirakhor (2016). Does trust contribute to stock market development? *Economic Modelling 52*, 239–50.

Nott, D. and R. Kohn (2005). Adaptive sampling for Bayesian variable selection. *Biometrika 92*, 747–63.

Oberdabernig, D., S. Humer, and J. Crespo Cuaresma (2017). Democracy, geography and model uncertainty. *Scottish Journal of Political Economy*, forthcoming.

Onorante, L. and A. E. Raftery (2016). Dynamic model averaging in large model spaces using dynamic Occam's window. *European Economic Review 81*, 2–14. Special issue on "Model Uncertainty in Economics".

Ouysse, R. (2016). Bayesian model averaging and principal component regression forecasts in a data rich environment. *International Journal of Forecasting 32*, 763–87.

Papageorgiou, C. (2011). How to use interaction terms in BMA: reply to Crespo Cuaresma's comment on Masanjala and Papageorgiou (2008). *Journal of Applied Econometrics 26*, 1048–50.

Pelster, M. and J. Vilsmeier (2016). The determinants of CDS spreads: evidence from the model space. Discussion Paper 43/2016, Deutsche Bundesbank.

Pérez, J. and J. Berger (2002). Expected-posterior prior distributions for model selection. *Biometrika 89*, 491–511.

Perrakis, K., D. Fouskakis, and I. Ntzoufras (2015). Variations of the power-conditional-expected-posterior prior for Bayesian variable selection in generalized linear models. ArXiv 1508.00793, Athens University of Economics and Business.

Pham, T. (2017). Impacts of globalization on the informal sector: Empirical evidence from developing countries. *Economic Modelling 62*, 207–18.

Philips, A. (2016). Seeing the forest through the trees: a meta-analysis of political budget cycles. *Public Choice 168*, 313–41.

Piironen, J. and A. Vehtari (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing 27*, 711–35.

Piribauer, P. (2016). Heterogeneity in spatial growth clusters. *Empirical Economics 51*, 659–80.

Piribauer, P. and J. Crespo Cuaresma (2016). Bayesian variable selection in spatial autoregressive models. *Spatial Economic Analysis 11*, 457–79.

Puy, D. (2016). Mutual funds flows and the geography of contagion. *Journal of International Money and Finance 60*, 73–93.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology 25*, 111–63.

Raftery, A., J. Hoeting, C. Volinsky, I. Painter, and K. Yeung (2010). Bayesian model averaging. R package vs. 3.13. `http://CRAN.R-project.org/package=BMA`.

Raftery, A., D. Madigan, and J. Hoeting (1997a). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 179–91.

Raftery, A., D. Madigan, and J. Hoeting (1997b). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 179–191.

Raftery, A. E., M. Kárný, and P. Ettler (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics 52*, 52–66.

Ribeiro, P. (2017). Selecting exchange rate fundamentals by bootstrap. *International Journal of Forecasting 33*, 894–914.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation* (2nd ed.). Springer.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods*. New York: Springer.

Rockey, J. and J. Temple (2016). Growth econometrics for agnostics and true believers. *European Economic Review 81*, 86–102. Special issue on "Model Uncertainty in Economics".

Rose, A. and M. Spiegel (2011). Cross-country causes and consequences of the crisis: An update. *European Economic Review 55*, 309–24.

Rossell, D., J. D. Cook, D. Telesca, and P. Roebuck (2014). mombf: Moment and inverse moment Bayes factors, R package version 1.5.9. `http://CRAN.R-project.org/package=mombf`.

Rossell, D. and D. Telesca (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association 112*, 254–65.

Rossi, P. E., G. Allenby, and R. McCulloch (2006). *Bayesian Statistics and Marketing*. New York: Wiley.

Russell, N., T. Murphy, and A. Raftery (2015). Bayesian model averaging in model-based clustering and density estimation. Technical Report arXiv: 1506.09035, University of Washington.

Sabanés Bové, D. (2011). hyper-g priors for GAM selection, R package. `https://r-forge.r-project.org/projects/hypergsplines/`.

Sabanés Bové, D. and L. Held (2011a). Bayesian fractional polynomials. *Statistics and Computing 29*, 309–24.

Sabanés Bové, D. and L. Held (2011b). Hyper-g priors for generalized linear models. *Bayesian Analysis 6*, 387–410.

Sabanés Bové, D., L. Held, and G. Kauermann (2015). Objective Bayesian model selection in generalised additive models with penalised splines. *Journal of Computational and Graphical Statistics 24*, 394–415.

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

Sala-i-Martin, X., G. Doppelhofer, and R. Miller (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review 94*, 813–35.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–4.

Scott, J. and J. Berger (2010). Bayes and empirical Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics 38*, 2587–619.

Shi, J. (2016). Bayesian model averaging under regime switching with application to cyclical macro variable forecasting. *Journal of Forecasting 35*, 250–62.

Som, A., C. Hans, and S. MacEachern (2015). Bayesian modeling with mixtures of block g priors. Technical report, Dept. of Statistics, Ohio State University.

Sparks, D., K. Khare, and M. Ghosh (2015). Necessary and sufficient conditions for high-dimensional posterior consistency under g-priors. *Bayesian Analysis 10*, 627–664.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B 64*, 583–640.

Steel, M. (2011). Bayesian model averaging and forecasting. *Bulletin of E.U. and U.S. Inflation and Macroeconomic Analysis 200*, 30–41.

Stock, J. and M. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*, 405–30.

Stock, J. and M. Watson (2006). Forecasting with many predictors. In C. G. G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, pp. 515–54. Elsevier.

Strachan, R. and H. van Dijk (2013). Evidence on features of a DSGE business cycle model from Bayesian model averaging. *International Economic Review 54*, 385–402.

Strachan, R. W. (2009). Comment on 'jointness of growth determinants' by Gernot Doppelhofer and Melvyn Weeks. *Journal of Applied Econometrics 24*, 245–47.

Strawderman, W. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics 42*, 385–8.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–88.

Tobias, J. and M. Li (2004). Returns to schooling and Bayesian model averaging; a union of two literatures. *Journal of Economic Surveys 18*, 153–80.

Traczynski, J. (2017). Firm default prediction: A Bayesian model-averaging approach. *Journal Of Financial And Quantitative Analysis 52*, 1211–45.

Tsangarides, C. G. (2004). A Bayesian approach to model uncertainty. Working Paper 04/68, IMF.

Vallejos, C. and M. Steel (2017). Bayesian survival modelling of university outcomes. *Journal of the Royal Statistical Society, A 180*, 613–31.

van den Broeck, J., G. Koop, J. Osiewalski, and M. Steel (1994). Stochastic frontier models: a Bayesian perspective. *Journal of Econometrics 61*, 273–303.

van der Maas, J. (2014). Forecasting inflation using time-varying Bayesian model averaging. *Statistica Neerlandica 68*, 149–82.

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

Vašíček, B., D. Žigraiová, M. Hoeberichts, R. Vermeulen, K. Šmídková, and J. de Haan (2017). Leading indicators of financial stress: New evidence. *Journal of Financial Stability 28*, 240–57.

Villa, C. and J. Lee (2016). Model prior distribution for variable selection in linear regression models. Discussion paper, University of Kent.

Villa, C. and S. Walker (2015). An objective Bayesian criterion to determine model prior probabilities. *Scandinavian Journal of Statistics 42*, 947–66.

Volinsky, C., D. Madigan, A. Raftery, and R. Kronmal (1997). Bayesian model averaging in proportional hazards model: Predicting the risk of a stroke. *Applied Statistics 46*, 443–8.

Wagner, M. and J. Hlouskova (2015). Growth regressions, principal components augmented regressions and frequentist model averaging. *Jahrbücher fur Nationalökonomie und Statistik 235*, 642–62.

Wallis, K. (2005). Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics 67*, 983–94.

Wang, H., X. Zhang, and G. Zou (2009). Frequentist model averaging estimation: A review. *Journal of Systems Science and Complexity 22*, 732–48.

Wang, M. (2017). Mixtures of g-priors for analysis of variance models with a diverging number of parameters. *Bayesian Analysis 12*, 511–32.

Wang, M. and Y. Maruyama (2016). Consistency of Bayes factor for nonnested model selection when the model dimension grows. *Bernoulli 22*, 2080–100.

Watson, P. and S. Deller (2017). Economic diversity, unemployment and the Great Recession. *The Quarterly Review of Economics and Finance 64*, 1–11.

Wei, Y. and Y. Cao (2017). Forecasting house prices using dynamic model averaging approach: Evidence from China. *Economic Modelling 61*, 147–55.

Wölfel, K. and C. Weber (2017). Searching for the Fed's reaction function. *Empirical Economics 52*, 191–227.

Womack, A., C. Fuentes, and D. Taylor-Rodriguez (2015). Model space priors for objective sparse Bayesian regression. arXiv 1511.04745v1.

Wright, J. (2008). Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics 146*, 329–41.

Wu, H., M. Ferreira, and M. Gompper (2016). Consistency of hyper-g-prior-based Bayesian variable selection for generalized linear models. *Brazilian Journal of Probability and Statistics 30*, 691–709.

Xiang, R., M. Ghosh, and K. Khare (2016). Consistency of Bayes factors under hyper g-priors with growing model size. *Journal of Statistical Planning and Inference 173*, 64 – 86.

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

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Amsterdam: North-Holland, pp. 233–43.

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

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

Zhang, H., X. Huang, J. Gan, W. Karmaus, and T. Sabo-Attwood (2016). A two-component g-prior for variable selection. *Bayesian Analysis 11*, 353–80.

Zhang, X., D. Yu, G. Zou, and H. Liang (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed- effects models. *Journal of the American Statistical Association 111*, 1775–90.

Zigraiova, D. and T. Havranek (2016). Bank competition and financial stability: Much ado about nothing? *Journal of Economic Surveys 30*, 944–81.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–29.