JOURNAL OF
ECONOMIC
SURVEYS

doi: 10.1111/joes.12044

MODEL AVERAGING IN ECONOMICS: AN OVERVIEW

Enrique Moral-Benito

Banco de España

Abstract. Standard practice in empirical research is based on two steps: first, researchers select a model from the space of all possible models; second, they proceed as if the selected model had generated the data. Therefore, uncertainty in the model selection step is typically ignored. Alternatively, model averaging accounts for this model uncertainty. In this paper, I review the literature on model averaging with special emphasis on its applications to economics. Finally, as an empirical illustration, I consider model averaging to examine the deterrent effect of capital punishment across states in the USA.

Keywords. Capital punishment; Model averaging; Model uncertainty

1. Introduction

The common practice in empirical research is based on selecting a single model after what amounts to a search in the space of all possible models. Then, researchers typically base their conclusions on this model acting as if the model chosen is the true model. However, this procedure tends to understate the real uncertainty and thus the conclusions might not be sufficiently conservative.[1]

Statistical models can be decomposed in two parts: the first one representing structural assumptions such as functional forms, control variables included or distributional choices for the residuals, and the second one representing parameters whose interpretation is specific to the imposed structural assumptions. Draper (1995) points out that 'even in controlled experiments and randomized sample surveys key aspects of the structure will usually be uncertain, and this is even more true with observational studies'.

Given the above, researcher's uncertainty about the value of the estimate of interest exists at distinct two levels. The first one is the uncertainty associated with the estimate conditional on a given model. This level of uncertainty is of course assessed in virtually every empirical study. What is not fully assessed is the uncertainty associated with the specification of the empirical model. It is typical for a given paper that the empirical specification is taken as essentially known; while some variations of a baseline model are often reported, standard empirical practice does not systematically account for the sensitivity of claims about the estimate of interest to model selection.

Depending on the context, candidate models to be selected might be substantially different in terms of functional form or distributional assumptions. However, the most common situation in economics refers to the uncertainty surrounding model selection among 2^q possible models when q variables are available for inclusion. This uncertainty is particularly relevant in open-ended economic applications in which the set of possible regressors can grow unwieldy because additional explanatory variables are compatible with each other.[2]

From a pure empirical viewpoint, model uncertainty represents a concern because estimates may well depend on the particular model considered. In the early 1980s, Leamer (1983) already pointed out the importance of the fragility of regression analysis to arbitrary decisions about the choice of control variables.

For the sake of intuition, imagine a situation in which there are many different candidate models for estimating the effect of X on Y. Facing this challenge, one can select a single model based on different criteria and then make inference based on that selected model ignoring the uncertainty surrounding the model selection process. The model selection literature has proposed different alternatives to carry out the selection step; the book by Claeskens and Hjort (2008) is an excellent reference.

An alternative strategy is to estimate all the candidate models and then compute a weighted average of all the estimates for the coefficient on X (i.e. the model averaging approach). We can then make inference based on the whole universe of candidate models. As a result, we consider not only the uncertainty associated to the parameter estimate conditional on a given model, but also the uncertainty of the parameter estimate across different models. In general, this approach leads us to wider confidence intervals for the estimated effect of X on Y with the hope that, in retrospect, researchers avoid noticing that their confidence bands were not sufficiently wide.

Frequentist Model Averaging (FMA) and Bayesian Model Averaging (BMA) are two different approaches to model averaging in the literature. Despite their similarities in spirit and objectives, both techniques differ in the approach to inference. Compared with the FMA approach, there has been a huge literature on the use of BMA in statistics and more recently in economics. Thus, the BMA toolkit is larger than that of FMA. However, the FMA approach is starting to receive a lot of attention over the last decade. This paper summarizes the state of the art in both approaches.

As empirical illustration of the model averaging methods surveyed in the paper, I consider the thorny debate on the deterrent effect of capital punishment. Model uncertainty plagues the empirical specifications considered in the literature estimating the effect of executions on crime rates (see Donohue and Wolfers, 2006). Therefore, I apply model averaging methods to US state-level data on murders, executions and a set of 16 control variables previously considered in the literature. Empirical estimates suggest that model uncertainty is large enough to preclude us from disentangling whether capital punishment has a deterrent effect.

On the other hand, given the raising interest on causal effects in economics over the last decades, the combination of model averaging and instrumental variables and panel data models is an interesting line of open research. The first steps in this direction have been taken over the last years (e.g. Durlauf *et al.* (2008) and Moral-Benito (2012a)). In this paper, I summarize some recent developments in this strand of the literature.

1.1 *Empirical Illustration: The Deterrent Effect of Capital Punishment*

The debate over capital punishment in the USA has continued for centuries, especially over the last decades, after the end of the moratorium by the Supreme Court.[3] Whether capital punishment deters murders has been extensively investigated in the literature since it represents a key issue in this debate (see Donohue and Wolfers (2006) for a recent overview).

The wide variation in execution rates across states during the post moratorium period serves as the empirical basis for these studies. In particular, cross-state regressions of the murder rate on the execution rate plus a set of control variables (such as economic and demographic indicators) are at the root of this literature. However, after three decades of research there is no consensus about the magnitude and significance of the deterrent effect of capital punishment. Model uncertainty in the specification of the empirical model (i.e. the choice of appropriate control variables) can explain, at least in part, the mixed findings in the literature (e.g. Cohen-Cole *et al.*, 2009).

To illustrate the importance of model uncertainty in this setting, I consider an empirical application to the state-level data in Donohue and Wolfers (2006). In particular, I regress state-level murder rates on the

Table 1. The Deterrent Effect of Capital Punishment.

	Dependent variable is the murder rate		
	(1)	(2)	(3)
Execution rate	0.19	−0.44	−11.49
	(0.02)	(−0.18)	(−2.72***)
Controls included	None	KLS (2003)	DS (2006)
N	51	51	51
R^2	0.01	0.68	0.85
Net lives saved per execution	−1.47	0.06	26.93
	(−0.07)	(0.01)	(2.62***)

Notes: KLS (2003) controls include prison deaths rate, prisoners per violent crime, prisoners per 100, 000 residents, per capita income, insured unemployment rate, percent of black and urban population, age distribution variables (fractions 0–24 and 25–44 year-olds) and infant mortality rates. DS (2006) controls include per capita income, unemployment rate, police employment, percent minority and age distribution variables (fractions 15–19 and 20–24 year-olds). State-level data refer to the year 2000. *t*-ratios are in parentheses. *, ** and *** indicate significance at the 10%, 5% and 1% levels, respectively.

execution rates using two different sets of control variables previously considered in two major studies in the deterrence literature. On the one hand, I include the controls in Dezhbakhsh and Shepard (2006), who find a strong deterrent effect from the death penalty. On the other hand, I also consider an alternative specification with the control variables suggested in Katz *et al.* (2003), who find that capital punishment does not represent a deterrent.

Both studies are based on the same choice–theoretic version of criminal behaviour advanced by Ehrlich (1975) in the capital punishment context. However, this theory does not provide strong guidance on how to construct a statistical model that maps theory to empirics. As a result, model uncertainty naturally arises in the specification of the empirical model of the deterrent effect of capital punishment.

Using the state-level data in Donohue and Wolfers (2006) for the year 2000, I estimate the following model[4]:

$$murder\ rate_i = \beta\ execution\ rate_i + \gamma x_i + u_i \tag{1}$$

where x_i refers to the vector of control variables for state i included in the regression, and β is the coefficient capturing the deterrent effect of capital punishment. From this coefficient we can also estimate the net lives saved by an additional execution as $NLS = \beta \frac{\text{pop}}{100{,}000} \frac{1}{\#\text{execut}}$ (see Donohue and Wolfers (2006)).

Table 1 presents the results of estimating equation (1). In addition to the naive specification in column (1), that is, with a constant and without control variables, I consider the two choices of x_i suggested by Katz *et al.* (2003) and Dezhbakhsh and Shepard (2006) as mentioned. The striking difference between β estimates in columns (2) and (3) is the basic illustration of model uncertainty in this setting.

In column (2) of Table 1, capital punishment does not have a significant deterrent effect, while column (3) presents a large and significant effect of death penalty on the murder rate, implying a net saving of around 26 lives from an additional execution. Estimates in both columns are based on exactly the same data set. The only difference is the set of control variables included in the regressions (see Table 1). Since *a priori* it is not clear which specification should be preferable, the researcher is left with two opposite findings and without enough guidance for selecting the preferred estimate. Moreover, one can also imagine thousands of alternative specifications given by different combinations of the control

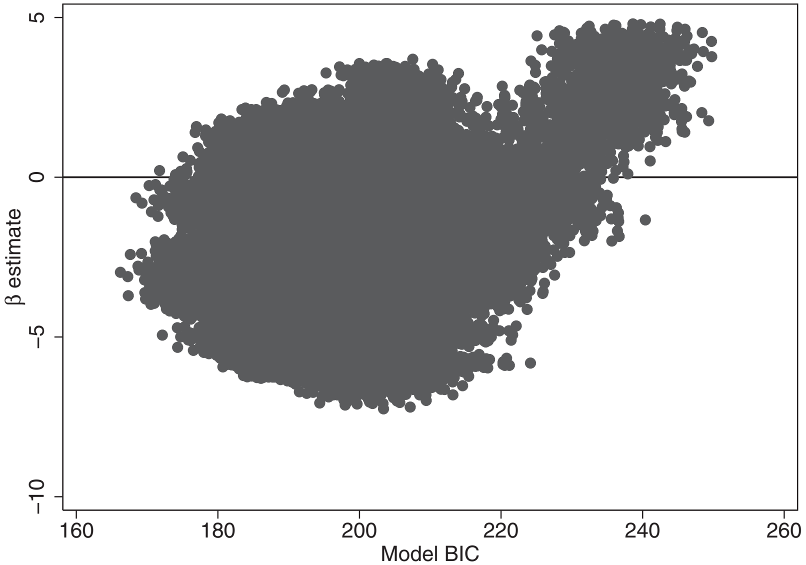

Figure 1. β Estimates from All Candidate Models.

This figure plots the β estimates corresponding to all the 65, 536 candidate models given by different combinations of control variables with the corresponding Bayesian Information Criterion (BIC) as a measure of goodness-of-fit for each specification. The execution rate and a constant are included in all cases.

variables considered in either Katz *et al.* (2003) or Dezhbakhsh and Shepard (2006) (and compiled together by Donohue and Wolfers (2006)).

Figure 1 presents the β estimates from the 2^{16} candidate models resulting from different combinations of the control variables in the x vector. It seems clear that model uncertainty is an issue in this particular application since the estimate of interest very much depends on the choice of control variables considered. Moreover, theory does not offer enough guidance to select a particular specification and, as illustrated by the Bayesian Information Criterion (BIC) in the x-axis of the graph, the best models in terms of goodness-of-fit are quite dispersed across different β estimates.

Throughout this paper, we discuss different alternatives to deal with such a situation from a model averaging perspective. Moreover, I examine the deterrent effect of capital punishment using different model averaging techniques in Section 4.

1.2 *A Historical Perspective on Model Combination*

As pointed out by Clemen (1989), Laplace (1818) considered combining regression coefficient estimates almost 200 years ago. In particular, he derived and compared the properties of two estimators, one being least squares and the other a kind of weighted median. Moreover, he also analysed the joint distribution of the two, and proposed a combining formula that resulted in a better estimator than either. Stigler (1973) presents a brief description of Laplace's work.

Aside from Laplace, other early treatments of combining multiple estimates came from the statistical literature. Edgerton and Kolbe (1936) propose to combine different estimates in such a way that the combining weights result from minimizing the sum of squares of the differences of the scores. Horst

(1938) derives a formula for combining multiple measures in which the criterion is obtaining maximum separation among the individual population members, and Halperin (1961) provided a minimum-squared-error combination of estimates. By the late 1970s, the idea of combining estimates was present, implicitly or explicitly, in several studies in the field of statistics (e.g. de Finetti (1972), Davis (1979) and Geisser and Eddy (1979)). More recently, Draper (1995) provides an assessment of the importance of model uncertainty in statistics and several alternatives to take it into account based on estimates combination.

In the forecasting literature, a flood of papers about combining different forecasts was generated in the 1960s and the 1970s since the influential papers by Barnard (1963) and Bates and Granger (1969). By that time, the idea of combining forecasts was well established in this literature, for example, Clemen (1989) surveyed over 200 studies from the late 1960s on the topic of forecast combination. Timmermann (2006) provides a good overview of recent advances in this literature.

The forecasting combination articles in the 1970s can be considered the predecessors of the current FMA literature. The FMA approach has started to receive attention over the last decade; see, for example, Hjort and Claeskens (2003) and Hansen (2007). This is so probably because the Frequentist approach to model uncertainty was traditionally focused on model selection rather than model averaging.

Geisser (1965), Roberts (1965) and Geisel (1973) appear to be the earliest Bayesian approaches to combining estimates. However, Leamer (1978) presents the first comprehensive description of the basic paradigm for BMA and therefore, it is typically cited as the seminal paper in the BMA literature. With a few exceptions such as Moulton (1991), the BMA approach was basically ignored in economic applications until the late 1990s and 2000s, when the 'BMA revolution' in economics took place.[5] This is so because more powerful computers and dramatic increases in numerical methods such as Monte Carlo Markov-Chain Model Composition (MC^3) allow applied researchers to overcome the troubles related to implementing BMA by exploring large model spaces in sensible ways. The state of research in the field during the nineties was summarized in Hoeting *et al.* (1999); two influential articles considering BMA in economics are Raftery (1995) and Fernández *et al.* (2001b).[6]

2. Bayesian Model Averaging

2.1 *Estimation and Inference with BMA*

For the sake of illustration, let us consider the case of a normal linear regression model in which model uncertainty comes from the selection of regressors to include in the right-hand side:

$$\begin{aligned} y &= X\beta + \epsilon \qquad (2) \\ \epsilon &\sim N(0, \sigma^2 I_T) \end{aligned}$$

where $y = (y_1, \ldots, y_T)'$ and ϵ are $T \times 1$ vectors of the dependent variable and the random shocks, respectively. X is a $T \times q$ matrix of regressors that may or may not be included in the model, and β $(q \times 1)$ contain the parameters to be estimated. If we set some components of $\beta = (\beta_1, \beta_2, \ldots, \beta_q)'$ to be zeros, there are a total of 2^q candidate models to be estimated – indexed by M_j for $j = 1, \ldots, 2^q$ – which all seek to explain y – the data.

Each model M_j depends upon parameters β^j. In cases where many models are being entertained, it is important to be explicit about which model is under consideration. Hence, following the Bayesian logic, the posterior for the parameters calculated using M_j is written as:

$$g\left(\beta^j | y, M_j\right) = \frac{f\left(y|\beta^j, M_j\right) g\left(\beta^j | M_j\right)}{f\left(y|M_j\right)} \qquad (3)$$

and the notation makes clear that we now have a posterior $g\left(\beta^j|y, M_j\right)$, a likelihood $f\left(y|\beta^j, M_j\right)$ and a prior $g\left(\beta^j|M_j\right)$ for each model.

Given a prior model probability $P\left(M_j\right)$ we can calculate the posterior model probability using Bayes Rule as:

$$P\left(M_j|y\right) = \frac{f\left(y|M_j\right)P\left(M_j\right)}{f\left(y\right)} \tag{4}$$

According to equations (3) and (4), it is now clear that we need to elicit priors for the parameters of each model and for the model probability itself.[7]

Following Leamer (1978) we can consider β a function of β^j for each $j = 1, \ldots, 2^q$ (i.e. $\beta(\beta^j)$) and then calculate the posterior density of the parameters for all the models under consideration by the law of total probability:

$$g\left(\beta|y\right) = \sum_{j=1}^{2^q} P\left(M_j|y\right) g\left(\beta|y, M_j\right) \tag{5}$$

Therefore, the full posterior distribution of β is a weighted average of its posterior distributions under each of the models, where the weights are given by $P\left(M_j|y\right)$. When applying BMA according to equation (5), both estimation and inference come naturally together from the posterior distribution. This posterior distribution provides inference about β that takes full account of model uncertainty.

One might also be interested in point estimates and their associated variances. One common procedure is to take expectations across (5):

$$E\left(\beta|y\right) = \sum_{j=1}^{2^q} P\left(M_j|y\right) E\left(\beta|y, M_j\right) \tag{6}$$

with associated posterior variance:

$$\begin{aligned} V\left(\beta|y\right) = & \sum_{j=1}^{2^q} P\left(M_j|y\right) V\left(\beta|y, M_j\right) + \\ & + \sum_{j=1}^{2^q} P\left(M_j|y\right) \left(E\left(\beta|y, M_j\right) - E\left(\beta|y\right)\right)^2 \end{aligned} \tag{7}$$

The posterior variance in (7) incorporates not only the weighted average of the estimated variances of the individual models but also the weighted variance in estimates of the coefficients β across different models.

Within the BMA approach, we can also compute the posterior inclusion probability (PIP) for a given variable. This PIP is calculated as the sum of the posterior model probabilities for all models including that variable.

Implementing BMA can be difficult because of two reasons: (i) two types of priors (on parameters and models) need to be elicited for many models, and this can be a complicated task. (ii) the number of models under consideration – 2^q – is often so large that the computational burden of BMA can be prohibitive. In the next sections I present some of the remedies proposed in the literature to these challenges.

2.2 *Priors on the Parameter Space*

Bayesian model averaging requires prior distributions for the unknown parameters under the various models M_j for $j = 1, \ldots, 2^q$, even when there is not enough prior knowledge to elicit them. In situations in which the researcher has little or no information about the unknown parameters, improper priors (i.e. priors that do not integrate to one) have been commonly employed as representations of this ignorance.

Given the difficulties of working with non-informative improper priors in the BMA framework, mainstream priors in BMA include hierarchical prior structures involving improper priors for the common (to all models) parameters, and proper priors for the remaining parameters. However, the proper priors typically considered are aimed to be partly non-informative (see Fernández *et al.*, 2001a).

2.2.1 *Improper Priors*

Since improper priors often lead to proper posteriors, the use of improper priors is not always a problem as long as the analysis is based on a single model. However, when comparing different models (as we do within the BMA approach), improper priors on all of the parameters result in ill-defined Bayes factors that depend on the ratio of two unspecified constants. Also, the resulting posterior model probabilities prefer (with probability one) the smaller model regardless of the information in the data (see Bartlett, 1957).

To circumvent the difficulty of using improper priors for model comparison various approaches have been advocated. One possibility is the 'imaginary training sample device' of Spiegelhalter and Smith (1982). This basically consists of assigning a value of 1 to the Bayes factor between models i and j based on a small subsample of the data. This yields a value for the ratio of constants. Alternatively, as suggested by Lempers (1971), one can set aside part of the data to use as a training sample which is combined with the improper prior distribution to produce a proper posterior distribution. Such posterior can be considered as a proper prior, and the Bayes factor is then computed from the remainder of the data.[8]

Given the difficulties of implementing these methods to many competing models simultaneously, BMA is typically based on proper and partly non-informative priors.

2.2.2 *Zellner's g Priors*

Given the normal regression framework, the bulk of the BMA literature favours the natural-conjugate approach, which puts a conditionally normal prior on coefficients β^j. Virtually all BMA studies use a conditional prior for the j-th model's parameters ($\beta^j|\sigma^2$) with zero mean and the variance proposed by Zellner (1986), that is, a prior covariance given by $g(X_j'X_j)^{-1}$. This prior variance is proportional to the posterior covariance arising from the sample ($(X_j'X_j)^{-1}$) with the scalar g determining how much importance is attributed to the prior beliefs of the researcher. The conditional prior on β^j is then:

$$\beta^j|\sigma^2, M_j, g \sim N(0, \sigma^2 g(X_j'X_j)^{-1}) \tag{8}$$

Moreover, the variance parameter σ is common to all the models under consideration, so an improper prior is not problematic. The most common approach is the uninformative prior first considered by Smith and Kohn (1996): $p(\sigma) \propto \sigma^{-1}$. If a constant term ($\alpha$) is included in all the models, we can also set the prior $p(\alpha) \propto 1$.

The popularity of this prior structure is due to two factors: (i) it has closed-form solutions for the posterior distributions that drastically reduce the computational burden, and (ii) it only requires the elicitation of one hyperparameter, the scalar g.

Since Smith and Kohn (1996) considered $g = 100$, many different options for choosing g have been proposed in the literature (see e.g. Fernández *et al.* (2001a)). However, I summarize here three of the most popular alternatives:

1. Unit Information Prior (g-UIP): proposed by Kass and Wasserman (1995), it corresponds to taking $g = N$, and it leads to Bayes factors that behave like the BIC. Therefore it is possible to combine

Frequentist OLS or MLE for estimation with the Schwarz approximation to the marginal likelihood for averaging with a Bayesian justification (see e.g. Raftery (1995) or Sala-i-Martin *et al.* (2004)).

2. Risk Inflation Criterion (g-RIC): recommended by Foster and George (1994), it implies setting $g = q^2$.
3. Benchmark Prior: After a thorough study, Fernández *et al.* (2001a) determined this combination of the g-UIP and g-RIC priors to perform best with respect to predictive performance. It matches with $g = \max(N, q^2)$.

2.2.3 *Laplace Priors*

Let us construct a partition of the X matrix such that we can rewrite (2) as follows:

$$\begin{aligned} y &= X_1\gamma + X_2\delta + \epsilon \\ \epsilon &\sim N(0, \sigma^2 I_N) \end{aligned} \tag{9}$$

where γ and δ are the new $q_1 \times 1$ and $q_2 \times 1$ parameter vectors with $q_1 + q_2 = q$.

Given this unrestricted model, we can determine which are the focus regressors (X_1) and which are the auxiliary (doubtful) regressors (X_2).[9] We can reparametrize the model in (9) replacing $X_2\delta = X_2^*\delta^*$, with $X_2^* = X_2 P \Pi^{-1/2}$ and $\delta^* = \Pi^{1/2} P'\delta$, where P is an orthogonal matrix and Π is a diagonal matrix such that $P'X_2' R_{X_1} X_2 P$ and $R_{X_1} = I - X_1(X_1'X_1)^{-1}X_1'$.

In this setting, Magnus *et al.* (2010) propose to consider an alternative prior structure that leads to the so-called weighted-average least squares (WALS) estimator. In particular, WALS use a Laplace distribution with zero mean for the independently and identically distributed elements of the transformed parameter vector $\eta = \delta^*/\sigma$, whose i-th element, η_i ($i = 1, \ldots, q_2$) is the population t-ratio on δ_i, the i-th element of δ. As pointed out by Magnus *et al.* (2010), ‘this choice of prior moments is based on our idea of ignorance as a situation where we do not know whether the theoretical t-ratio is larger or smaller than one in absolute value’.

The WALS estimator employs non-informative model-specific priors and drastically reduces the computational burden of standard BMA being proportional to q_2 (or q) instead of 2^{q_2} (or 2^q). In contrast, WALS does not provide either Bayesian posterior distributions or posterior inclusion probabilities as a measure of robustness.

2.2.4 *Empirical Bayes (EB) Priors*

In the context of BMA, EB approaches estimate the hyperparameter g from the data rather than pre-select g *a priori*. In fact, many Bayesians are critical of this approach on the grounds that it does not correspond to a formal Bayesian procedure.[10]

Two common EB alternatives are available in the literature, namely, the local EB approach developed by Hansen and Yu (2001), and the global EB approach suggested by George and Foster (2000) and Clyde and George (2000).

The local EB approach estimates a separate g for each model. Since Zellner’s g priors allow obtaining closed-form expressions of all model-specific marginal likelihoods, local EB uses the marginal likelihood of each model M_j for estimating g_j ($j = 1, \ldots, 2^q$). In particular, the local EB estimate of g_j is the maximum likelihood estimate constrained to be non-negative, which is given by:

$$\hat{g}_j^{EBL} = \max\{F_j, 0\} \tag{10}$$

where F_j is the F statistic for testing $\beta_j = 0$ in model j (see Liang *et al.* (2008)).

Alternatively, the global EB approach estimates a single g for all the models under consideration. More concretely, $\hat{g}_j^{EBG}$ is the maximum likelihood estimate resulting from maximization of the marginal likelihood of the data, obtained by averaging over all model-specific marginal likelihoods. In contrast to the local EB approach, there is no closed-form solution for $\hat{g}_j^{EBG}$, and thus, numerical optimization routines are typically considered (e.g. George and Foster, 2000).

2.3 *Priors on the Model Space*

In order to implement any of the BMA strategies described, prior model probabilities ($P(M_j)$) must be assigned. This step might be considered as analogous to the choice of model weights in the Frequentist approach to model averaging (more on this later).

2.3.1 *Binomial Priors*

For the model size (Ξ), the most common prior structure in BMA research is the Binomial distribution. According to this priors, each variable is independently included (or not) in a model so that model size (Ξ) follows a Binomial distribution with probability of success ξ:

$$\Xi \sim \text{Bin}(q, \xi) \tag{11}$$

where q is the number of regressors considered and ξ is the prior inclusion probability for each variable.

Given the above, the prior probability of a model (M_j) with q_j regressors is given by:

$$P(M_j) = \xi^{q_j}(1-\xi)^{q-q_j} \tag{12}$$

One commonly used particular case of this prior structure is to assume that every model has the same *a priori* probability (i.e. the uniform prior on the model space). This uniform prior corresponds to the assumption that $\xi = 1/2$ so that $P(M_j) = 2^{-q}$.[11] Moreover, given that $E(\Xi) = q\xi$, we can fix different priors in terms of both the prior inclusion probability (ξ) or the prior expected model size ($E(\Xi)$).

2.3.2 *Binomial-Beta Priors*

Ley and Steel (2009) propose an alternative prior specification in which ξ is treated as random rather than fixed. The proposed hierarchical prior implies a substantial increase in prior uncertainty about model size (Ξ), and makes the choice of prior model probabilities much less critical.

In particular, their proposal is the following:

$$\Xi \sim \text{Bin}(q, \xi) \tag{13}$$

$$\xi \sim \text{Be}(a, b) \tag{14}$$

where $a, b > 0$ are hyper-parameters to be fixed by the researcher. The difference with respect to the Binomial priors is to make ξ random rather than fixed. Model size Ξ will now satisfy:

$$E(\Xi) = \frac{a}{a+b}q \tag{15}$$

The model size distribution generated in this way is the so-called Binomial–Beta distribution. Ley and Steel (2009) propose to fix $a = 1$ and $b = (q - E(\Xi))/E(\Xi)$ through equation (15), so we only need to specify $E(\Xi)$, the prior expected model size, as in the Binomial priors. However, sensitivity of the posteriors with Binomial–Beta priors is smaller than with the Binomial priors.

2.3.3 *Dilution Priors*

Both the Binomial and the Binomial–Beta priors have in common the implicit assumption that the probability of one regressor appears in the model is independent of the inclusion of others, whereas regressors are typically correlated. In fact, with this priors on model space, a researcher could arbitrarily increase (or reduce) the prior model probabilities across theories simply by including redundant proxy variables for some of these theories. This is the denominated dilution problem raised by George (1999).

To address this issue, Durlauf *et al.* (2008) introduce a version of George (1999) dilution priors that assigns probability to neighbourhoods of models. Moreover, this kind of dilution prior assigns uniform probability to neighbourhoods rather than models, and solves the dilution problem. Consider a given theory (or neighbourhood of models) (T) for which we have q_T proxies among the whole set of q regressors. For each possible combination of variables corresponding to theory T (C_T) we can assign the following prior probability:

$$P(C_T) = |R_{C_T}| \prod_{h=1}^{q_T} \xi^{\pi_h}(1-\xi)^{1-\pi_h} \tag{16}$$

where π_h is an indicator of whether or not variable h is included in the combination C_T and R_{C_T} is the correlation matrix for the set of variables included in C_T. Since the determinant of this correlation matrix $(|R_{C_T}|)$ goes to 1 when the set of variables are orthogonal and to 0 when the variables are collinear, these priors are designed to penalize models with many redundant variables. In practice, we assign the same probability to all the models included in the neighbourhood C_T and uniform probability to all the different neighbourhoods.

Despite its advantages regarding the dilution property, this prior structure requires agreement on which regressors are proxies for the same theories (i.e. it requires to define the model neighbourhoods) which is usually not within reach.

2.4 *Further Topics in BMA*

2.4.1 *Computational Aspects*

In theory, with the results described above we should be able to carry out BMA. However, in practice, the number of models under consideration (2^q) is often so big that makes it impossible to estimate every possible model. Accordingly, there are several algorithms developed in the literature which carry out BMA without evaluating every possible model.

One possible approach is the so-called Occam's Window proposed by Madigan and Raftery (1994). The basic idea of this technique is to exclude from the summation models that predict the data far less well than the best model, and models that receive less support than any of their simpler submodels. Therefore, using an appropriate search strategy (for instance the leaps and bounds algorithm by Furnival and Wilson (1974)) the number of models to be estimated is drastically reduced.

Another commonly used alternative, initially developed in Madigan and York (1995) is MC^3. Markov Chain Monte Carlo (MCMC) methods are common in Bayesian econometrics. MCMC algorithms in general take draws from the parameter space in order to simulate the posterior distribution of interest. However, they do not draw from every region of the parameter space, but focus on regions of high posterior probability. BMA considers the models as discrete random variables so that posterior simulators which draw from the model space instead of the parameter space can be derived. As MCMC in the parameter space, MC^3 takes draws from the model space focusing on models with high posterior model probability. Implementing and programming MC^3 is intuitive and it is not very complicated (see chapter 11 in Koop (2003))

2.4.2 *Jointness*

A relevant issue which arises in the BMA framework is whether different sets of regressors are substitutes or complements in the determination of the outcome. Accounting for these interdependencies among the regressors delivers more parsimonious models with minimally reduced explanatory power.[12] Ley and Steel (2007) and Doppelhofer and Weeks (2009) define *ex post* measures of dependence among explanatory variables that appear in linear regression models. The object of interest in both approaches is the measure of jointness (or interdependency) of two regressors X_i and X_j in the context of linear regressions. For instance, Ley and Steel (2007) propose two alternative measures:

$$J^*_{LS} = \frac{P(i \cap j)}{P(i) + P(j) - P(i \cap j)} \in [0, 1] \tag{17}$$

$$J_{LS} = \frac{P(i \cap j)}{P(i) + P(j) - 2P(i \cap j)} \in [0, \infty) \tag{18}$$

where $P(i \cap j)$ is the sum of the posterior probabilities of the regression models that contain both X_i and X_j, and $P(i)$ and $P(j)$ are the posterior inclusion probabilities of X_i and X_j, respectively.

2.4.3 *A Frequentist Approach to BMA?*

If we assume diffuse priors on the parameter space for any given sample size, or, if we have a large sample for any given prior on the parameter space we can write equation (6) as follows:[13]

$$E(\beta|y) = \sum_{j=1}^{2^q} P\left(M_j|y\right) E(\beta|y, M_j) = \sum_{j=1}^{2^q} P\left(M_j|y\right) \widehat{\beta}^j_{ML} \tag{19}$$

where $\widehat{\beta}^j_{ML}$ is the ML estimate for model j.

If one is interested in model averaged point estimates, we can use the Schwarz asymptotic approximation to the Bayes factor and uniform model priors so that:

$$P\left(M_j|y\right) = \frac{f(y|\widehat{\beta}_j, M_j)N^{\frac{-q_j}{2}}}{\sum_{i=1}^{2^q} f(y|\widehat{\beta}_i, M_i)N^{\frac{-q_i}{2}}} \tag{20}$$

where $f(y|\widehat{\beta}_j, M_j)$ is the maximized likelihood function for model j.

Comparing this expression with Frequentist model weights based on information criteria (see Section 3.3.1), and given the use of maximum likelihood estimates, this commonly used approach to BMA (e.g. Raftery, 1995; Sala-i-Martin *et al.*, 2004; Moral-Benito, 2012a) can be labelled as a Frequentist BMA method.

This approach was first proposed by Raftery (1995) in a general setting. Sala-i-Martin *et al.* (2004) popularized its use in economics averaging model-specific OLS estimates in the so-called Bayesian Averaging of Classical Estimates (BACE). Finally, Moral-Benito (2012a) generalized the use of this approach to panel data models in the denominated Bayesian Averaging of Maximum Likelihood Estimates (BAMLE).[14]

3. Frequentist Model Averaging

3.1 *Definition of FMA Estimators*

Let us take the linear model in matrix form to illustrate the definition of the FMA estimator:

$$y = \beta X_A + X_B\gamma + \epsilon \tag{21}$$

where y, X_A and ϵ are $T \times 1$ vectors of the dependent variable, the treatment variable of interest and the random shocks, respectively. X_B is a $T \times q$ matrix of doubtful control variables that may or may not be included in the model, and β and γ $(q \times 1)$ contain the parameters to be estimated. Despite we make this distinction between X_A and X_B for illustration purposes, FMA can easily handle situations in which we cannot make such a distinction. Finally, T represents the number of observations in the sample.

If we set some components of $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)'$ to be zeros, there are a total of 2^q candidate models to be estimated. Given the coefficient of interest is β, let $\hat{\beta}_M$ be the estimator of β under the candidate model M with $M \in \{M_1, M_2, \ldots, M_{2^q}\}$. The most common approach in applied research is to take the selected model as given and base the inference on this single estimate $\hat{\beta}_M$ while the actual estimator is:

$$\hat{\beta} = \begin{cases} \hat{\beta}_{M_1} & \text{if the first model is selected} \\ \hat{\beta}_{M_2} & \text{if the second model is selected} \\ \vdots & \vdots \\ \hat{\beta}_{M_{2^q}} & \text{if the} 2^q\text{-th model is selected} \end{cases} \tag{22}$$

We can also rewrite the above estimator as

$$\hat{\beta} = \sum_{j=1}^{2^q} \tilde{\omega}_{M_j}\hat{\beta}_{M_j} \tag{23}$$

where:

$$\tilde{\omega}_{M_j} = \begin{cases} 1 & \text{if the candidate model} M_j \text{is selected} \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

This estimator is the usual pre-test estimator that suffers from the previously commented drawbacks if model uncertainty (in the selection of the control variables for example) is present. Therefore, we consider the smoothed weights ω_{M_j} and accordingly, the FMA estimator is:

$$\hat{\beta}_{\text{FMA}} = \sum_{j=1}^{2^q} \omega_{M_j}\hat{\beta}_{M_j} \tag{25}$$

where $0 \leq \omega_{M_j} \leq 1$, and $\sum_{j=1}^{2^q} \omega_{M_j} = 1$. Such estimator is labelled as the FMA estimator of β which integrates both model selection and parameter estimation.

3.2 *FMA Inference*

Hjort and Claeskens (2003) studied the asymptotic properties of the FMA estimator with the form in equation (25). The main result is the obtaining of its asymptotic distribution:

$$\sqrt{N}\left(\hat{\beta}_{\text{FMA}} - \beta_{\text{true}}\right) \xrightarrow{d} \Lambda \tag{26}$$

where $\Lambda = \sum_{j=1}^{2^q} \omega_{M_j} \Lambda_j$ represents the asymptotic distribution of the FMA estimator as detailed in Claeskens and Hjort (2008).

However, inference based on this limiting distribution Λ still ignores the uncertainty involved in the model selection step since its variance is based on averaging the model specific variances. Therefore, confidence intervals constructed from $\hat{\beta}_{\text{FMA}}$ and the variance of Λ are expected to produce too optimistic inference and they might lead to misleading conclusions because the real coverage probability is lower than the intended level.

In response to this problem, Buckland *et al.* (1997) proposed an alternative approach to deal with this issue when constructing confidence intervals of FMA estimators. Their method takes the extra model uncertainty into account by including an extra term in the variance of the FMA estimator. In particular, the proposed formula for the estimated standard error of $\hat{\beta}_{\text{FMA}}$ is:

$$\widehat{SE}(\hat{\beta}_{\text{FMA}}) = \sum_{j=1}^{2^q} \omega_{M_j} \sqrt{\hat{\tau}_j^2/N + \hat{b}_j^2} \tag{27}$$

where $\hat{\tau}_j^2$ estimates the variance of Λ_j, and $\hat{b}_j = \hat{\beta}_{M_j} - \hat{\beta}_{\text{FMA}}$ captures the extra uncertainty associated with the variation of estimates across different models (note that this extra term is not present in the variance of Λ in (26)). This formula implies an estimated variance for the FMA estimator that closely resembles its Bayesian counterpart in equation (7). Note that we still have to replace the fixed weights in equations (25) and (27) by their estimates in order to apply FMA.

3.3 *Model Weights in FMA*

FMA estimators crucially depend on the weights selected for estimation. In the previous subsections the weights were taken as fixed, but it is important to remark here that different weights will result in different asymptotic properties of the corresponding FMA estimators.

3.3.1 *Weight Choice Based on Information Criteria*

Probably the most common approach to weight choice in FMA is the one based on different information criteria of the form:

$$I_j = -2\log(L_j) + \varphi_j \tag{28}$$

where L_j is the maximized likelihood function for the j-th model, and φ_j is a penalty term function of the number of parameters and/or the number of observations of model j (i.e. q_j).

In the spirit of likelihood ratio methods, Buckland *et al.* (1997) propose to use the following model weights:

$$\omega_{M_j} = \frac{\exp(-I_j/2)}{\sum_{h=1}^{2^q} \exp(-I_h/2)} \tag{29}$$

which are also normalized to sum to unity.

The penalty term $\varphi_j = 2q_j$ corresponds to the Akaike Information Criterion (AIC), being q_j the number of parameters in model j. Another possible choice is $\varphi_j = q_j \ln(N)$ that corresponds to the BIC. Given the use of BIC is also justified from a Bayesian viewpoint, this illustrates one clear similarity between BMA and FMA.

Information criteria such as the AIC and the BIC select one single best model regardless of the parameter of interest. However, there are situations in which one model is best for estimating one parameter, whereas another model is best for another parameter. Aware of this situation, Claeskens and Hjort (2003) propose to use the Focused Information Criterion (FIC) to select the best model, but depending on the parameter of interest. Of course, the FIC can naturally be employed as an alternative to construct FMA model weights.

3.3.2 *Weight Choice Based on Mallows' Criterion*

Hansen (2007) proposes to select the model weights in least-squares model averaging by minimizing the Mallows' criterion. Despite this criterion is similar to the AIC in the model selection spirit, the approach to calculate the weights in Hansen (2007) is different.

Hansen (2007) considers the following linear regression:

$$y_i = \sum_{j=1}^{\infty} \theta_j x_{ij} + \epsilon_i \tag{30}$$

together with $E(\epsilon_i|x_i) = 0$ and $E(\epsilon_i^2|x_i) = \sigma^2$, where $x_i = (x_{i1}, x_{i2}, \ldots)$.

Now consider the sequence of candidate models $j = 1, 2, \ldots$ seeking to approximate (30). The j-th model uses the first ϕ_j elements of x_i with $0 < \phi_1 < \phi_2 < \ldots$. Given the above, the j-th candidate model is:

$$y_i = \sum_{j=1}^{\phi_j} \theta_j x_{ij} + \epsilon_i \tag{31}$$

with corresponding approximating error $\sum_{j=\phi_j+1}^{\infty} \theta_j x_{ij}$. Let us rewrite (31) in matrix form:

$$Y = X_j \Theta_j + \epsilon \tag{32}$$

where Y and ϵ are $T \times 1$ vectors, X_j is a $T \times \phi_j$ matrix and Θ_j is a $\phi_j \times 1$ vector of parameters. Let $J = J(T) \leq T$ be the candidate model with the largest number of regressors, and $\lambda = (\lambda_1, \ldots, \lambda_J)'$ a weight vector in the unit simplex in $\mathbb{R}^J$ with $\lambda \in \mathbf{H}_T$.

The least-squares model averaging estimator of Θ_J can be defined as:

$$\hat{\Theta}_J(\lambda) = \sum_{j=1}^{J} \lambda_j \begin{pmatrix} \hat{\Theta}_j \\ 0 \end{pmatrix} \tag{33}$$

where $\hat{\Theta}_j$ represents the least-squares estimate of model j.

We are now ready to introduce the Mallows' criterion to be minimized in order to obtain the model weights:

$$\hat{\lambda} = \underset{\lambda \in \mathbf{H}_T}{\operatorname{argmin}}\ C_T(\lambda) \tag{34}$$

where:

$$C_T(\lambda) = (Y - X_J \hat{\Theta}_J(\lambda))'(Y - X_J \hat{\Theta}_J(\lambda)) + 2\sigma^2 \lambda' \Phi \tag{35}$$

with $\Phi = (\phi_1, \ldots, \phi_J)'$. Note that the Mallows' criterion $C_T(\lambda)$ is an unbiased estimate of the expected squared error plus a constant.

Furthermore, Hansen (2007) provides an optimality result of his Mallows Model Averaging (MMA) estimator. In particular, it states that the MMA estimator achieves the lowest possible squared error when we constrain the weight vector to the discrete set $\mathbf{H}_T$ (i.e. it is asymptotically optimal). However, it is important to mention that the optimality of MMA fails under heteroskedasticity.

In a situation of instrument uncertainty (i.e. many candidate instruments for a given set of endogenous variables), Kuersteiner and Okui (2010) propose to apply the MMA approach to the first stage of the

2SLS, LIML and Fuller estimators, and then use the average predicted value of the endogenous variables in the second stage. On the other hand, Hansen (2008) considers forecast combination based on MMA, i.e., selecting forecast weights by minimizing a Mallows criterion.

3.3.3 *Weight Choice Based on Cross-Validation Criterion*

In a recent paper, Hansen and Racine (2012) propose how to optimally average across non-nested and heteroskedastic models. In particular, they suggest to select the weights of the least-squares model averaging estimator by minimizing a deleted-1 cross-validation criterion, so that the approach is labelled as Jackknife Model Averaging (JMA). In comparison with MMA, JMA (and its optimality result) is appropriate for more general linear models (i.e. random errors may have heteroskedastic variances, and the candidate models are allowed to be non-nested). Aside from these two points, the setup is the same as for the MMA estimator in (30).

Let us further define $\mu_i = E(y_i|\epsilon_i)$ so that:

$$\mu_i = \sum_{j=1}^{\infty} \theta_j x_{ij} \tag{36}$$

The jackknife version of the model averaging estimator of μ is:

$$\hat{\mu}(\lambda) = \sum_{j=1}^{J} \lambda_j \hat{P}_j Y = \hat{P}(\lambda) Y \tag{37}$$

where $\hat{P}_j = \hat{D}_j(P_j - I_T) + I_T$, $P_j = X_j(X_j' X_j)^{-1} X_j'$ is the projection matrix under the j-th candidate model, $\hat{D}_j$ is the $T \times T$ diagonal matrix with the i-th diagonal element being $(1 - h_{ii}^j)^{-1}$, $h_{ii}^j = X_{j,i}(X_j' X_j)^{-1} X_{j,i}'$ and $X_{j,i}$ is the i-th row of X_j. The deleted-1 cross-validation criterion is defined as:

$$CV(\lambda) = (Y - \hat{\mu}(\lambda))'(Y - \hat{\mu}(\lambda)) \tag{38}$$

which represents a quadratic form on the weights and it resembles the expected squared error.

Finally, the JMA estimator is $\hat{\mu}(\hat{\lambda}^*)$ with weights given by:

$$\hat{\lambda}^* = \underset{\lambda \in \mathbf{H}_T}{argmin}\, CV(\lambda) \tag{39}$$

Moreover, there is also a theorem that builds the asymptotic optimality of the JMA estimator in the sense of achieving the lowest possible expected squared error. Hansen and Racine (2012) also conduct Monte Carlo simulations showing that JMA can achieve significant efficiency gains over existing model selection and averaging methods in the presence of heteroskedasticity.

4. Model Averaging and The Deterrent Effect of Capital Punishment

As illustrated in the Introduction, model uncertainty plagues the empirical literature on the deterrent effect of capital punishment. In this section, I formally incorporate model uncertainty into the analysis using model averaging techniques. The aim of this exercise is to illustrate the usefulness and implementation details of model averaging. Important aspects such as endogeneity of the regressors and state-specific unobserved heterogeneity are not considered in this section. Cohen-Cole *et al.* (2009) provide an in-depth analysis of the deterrence literature accounting for model uncertainty.

The generic specification in equation (1) is the basis for the model space construction. In particular, I consider all possible combinations of the 16 candidate control variables that can be included in the vector x_i. Our variable of interest, the execution rate and the constant term are included in all specifications. This results in $2^{16} = 65,536$ candidate models to be combined within the model averaging framework.[15]

4.1 *Data*

All the data used in the exercise come from Donohue and Wolfers (2006), who compile and update the state-level data sets in Katz *et al.* (2003) and Dezhbakhsh and Shepard (2006). While the original data set contains panel information for 51 states in the USA over the period 1960–2000, I only use a cross section of states in the year 2000 for the sake of simplicity.

The dependent variable is the murder rate, that is, annual homicides per 100,000 residents; the deterrent variable of interest is the execution rate, that is, number of executions per 1000 prisoners. As for the remaining 16 regressors, I include all the candidate control variables considered by Katz *et al.* (2003) and Dezhbakhsh and Shepard (2006). They include demographic controls such as the percentages of population age 15–19, age 20–24, age 0–24 and age 25–44, the percentage of a state's population that is black, the percentage of population residing in urban areas, and infant mortality rates. The economic controls are real state per capita income, the unemployment rate and the insured unemployment rate.

To control for the severity of a state's criminal justice system the number of prisoners per violent crime and the ratio of prisoners to state population are also considered (see Katz *et al.*, 2003). As additional deterrent factors, the data include full-time state police employees (Dezhbakhsh and Shepard, 2006) and the number of prison deaths per 1000 prisoners (Katz *et al.*, 2003). Finally, the potential effects on murder rates driven by changes in other violent crimes are proxied by the rates of aggravated assaults and robberies (Dezhbakhsh and Shepard, 2006). More details on the variables and their sources can be found in Katz *et al.* (2003) and Dezhbakhsh and Shepard (2006). Also, the data set (compiled by Donohue and Wolfers (2006)) is available at http://bpp.wharton.upenn.edu/jwolfers/DeathPenalty.shtml.

4.2 *Results*

In order to incorporate model uncertainty into the estimation of the deterrent effect of capital punishment, I reexamine the regressions in Table 1 but employing the model averaging techniques described in Sections 2 and 3. This allows obtaining estimates of the deterrent effect that do not depend on the particular specification selected.

With respect to the Bayesian perspective, I employ three different prior schemes on the parameter space: the benchmark g-prior (BMA-g, Fernández *et al.* (2001a)), the Laplace prior (WALS, Magnus *et al.* (2010)) and the unit information g-prior (BMA-bic, Kass and Wasserman (1995)). In all the three cases I assume uniform priors on the model space. Turning to the Frequentist approach to model averaging, I consider three alternative weighting schemes: weights based on the AIC (FMA-aic), weights based on the Mallows' criterion (MMA, Hansen (2007)) and model weights based on cross-validation criteria (JMA, Hansen and Racine (2012)).

Table 2 presents the results. Estimates of the deterrent coefficient on the execution rate are very similar across the six model averaging schemes considered. However, model uncertainty renders estimations of the deterrent effect very imprecise as illustrated by the model averaging standard errors. Despite the model-averaged point estimate of net lives saved from an additional execution is positive in all columns, the associated standard error precludes us from concluding that the effect is statistically significant.[16]

With respect to the different control variables considered, only the percentage of a state's population that is black appears as robustly associated with murder rates across the states in our sample. For all the other controls, model uncertainty is large enough to hamper consensus on the regressors robustly related

Table 2. The Deterrent Effect of Capital Punishment Via Model Averaging.

	BMA-g (1)	WALS (2)	BMA-bic (3)	FMA-aic (4)	MMA (5)	JMA (6)
Execution rate	−2.477	−1.786	−2.652	−2.701	−2.712	−2.717
	(2.357)	(2.264)	(2.343)	(2.331)	(2.327)	(2.303)
Infant mortality rate	0.000	0.003	0.000	0.001	0.001	0.001
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Prison deaths rate	0.002	0.035	0.007	0.020	0.018	0.019
	(0.058)	(0.211)	(0.088)	(0.084)	(0.076)	(0.088)
Prisoners per 100,000 residents	0.007	0.005	0.007	0.007	0.007	0.007
	(0.005)	(0.003)	(0.005)	(0.004)	(0.004)	(0.004)
Prisoners per violent crime	−5.716	−4.072	−5.676	−5.783	−5.783	−5.946
	(5.270)	(3.106)	(4.894)	(4.275)	(4.398)	(4.280)
Fraction black	9.604	6.898	9.415	9.171	9.404	9.540
	(4.189)	(3.582)	(4.097)	(3.957)	(3.777)	(3.647)
Fraction urban	0.067	1.959	0.176	0.517	0.583	0.648
	(1.201)	(2.769)	(1.633)	(1.621)	(1.585)	(1.754)
Fraction 25–44 year-olds	2.488	23.296	5.401	11.209	12.553	12.061
	(10.382)	(20.588)	(14.417)	(17.490)	(18.232)	(17.818)
Fraction 0–24 year-olds	0.700	5.849	1.304	2.746	2.981	2.925
	(4.541)	(23.022)	(7.364)	(8.585)	(8.039)	(8.757)
Insured unemployment rate	−0.036	−0.285	−0.088	−0.182	−0.198	−0.202
	(0.167)	(0.372)	(0.255)	(0.293)	(0.305)	(0.308)
Unemployment rate	0.032	0.190	0.062	0.097	0.100	0.105
	(0.150)	(0.342)	(0.214)	(0.214)	(0.214)	(0.220)
Robbery rate	0.009	0.006	0.009	0.009	0.008	0.008
	(0.010)	(0.007)	(0.009)	(0.008)	(0.008)	(0.008)
Aggravated assault rate	0.000	0.001	0.000	0.001	0.001	0.001
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Per capita income	−0.016	−0.026	−0.020	−0.024	−0.025	−0.025
	(0.017)	(0.019)	(0.018)	(0.018)	(0.018)	(0.018)
Fraction 15-19 year-olds	2.513	43.535	4.555	12.870	12.420	13.501
	(18.257)	(89.313)	(28.666)	(35.387)	(31.376)	(35.575)
Fraction 20-24 year-olds	−1.519	−55.921	−6.671	−23.738	−25.933	−26.749
	(14.492)	(58.718)	(27.002)	(40.356)	(41.718)	(41.936)
Full-time police employees	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Net lives saved	5.02	3.34	5.44	5.56	5.59	5.60
	(4.73)	(4.50)	(4.69)	(4.66)	(4.65)	(4.60)

Notes: Dependent variable is the murder rate. BMA-g refers to BMA with benchmark g-priors on the parameter space (Section 2.2.1). WALS refers to the Laplace priors on the parameter space as described in Section 2.2.2. BMA-bic considers the g-UIP priors for the parameters, that is the Bayesian information criterion, as discussed in Sections 2.2.1 and 2.4.2. FMA-aic refers to FMA combined with model-specific weights based on the Akaike information criterion. MMA and JMA refer to Mallows model averaging (Section 3.3.2) and Jackknife model averaging (3.3.3), respectively. In columns (1)–(3) a uniform prior over the model space is assumed. Model averaging standard errors are in parentheses.

to murder rates. It is crucial to recognize that this analysis focuses on a particular specification only exploiting cross-state variation in the year 2000, and issues such as reverse causality are not properly accounted for.

All in all, one should interpret these results with caution, and only as an illustration of the usefulness of the model averaging techniques summarized in this paper.

The fact that the same conclusions emerge from all the six model averaging schemes is reassuring. Despite one must decide on several aspects when employing model averaging (priors on parameters, priors on models, combination weights, etc.), the key advantage of model averaging is present in all of them. Incorporating parameter estimate uncertainty across different models allows making inference that does not depend in any particular model.

5. Predictive Model Averaging

Forecasting (or prediction) is the process of making statements about events whose actual outcomes have not yet been observed. The most common forecasting situation involves the estimation of some variable of interest at some specified future date. It is also very common that the researcher has under consideration not a single forecast (or model) but several alternative forecasts (or models). Despite techniques for combining point forecasts have been in the toolkit of forecasters for many years, the literature on how to optimally combine probabilistic forecasts has burgeoned in recent years.

5.1 *Point Forecasts Combination*

Since Bates and Granger (1969), the idea that combinations of point forecasts outperform any individual forecast is part of the folklore of economic forecasting. When several point forecasts are available, the simplest combination is given by equal weights to each single forecast; this combination has proven to be hard to bit in practice giving rise to the forecast combination puzzle (e.g. Smith and Wallis, 2009).

In theory, if the available forecasts are uncorrelated, the optimal weights under mean squared loss should be inversely proportional to each forecast's variance. Bates and Granger (1969) suggested using empirical weights based on out-of-sample forecast variances. Granger and Ramanathan (1984) proposed to regress the actual value on the available forecasts and use the coefficient estimates (constrained to be non-negative and sum to one) as weights.[17]

Handling structural breaks remains a challenge to forecasters, and combination schemes have proven useful in this respect. In the presence of structural breaks, Pesaran and Timmermann (2005, 2007) find that the estimation window matters to out-of-sample forecasting performance. The same authors suggest approaches based on combinations of forecasts generated under different estimation windows (but the same model) rather than under different models. Moreover, they illustrate that these combination schemes outperform forecasts from individual models.

5.2 *Probabilistic Forecasts Combination*

Probabilistic (or density) forecasts provide the researcher with the full predictive probability density function (pdf) as opposed to a single moment of this pdf. As a result, probabilistic forecasts allow obtaining measures of uncertainty surrounding a 'central tendency' (the point forecast) that can enhance the usefulness of the forecast (e.g. Gneiting, 2008). Along these lines, Granger *et al.* (1989) consider combinations of confidence intervals or quantiles.

The consideration of probabilistic forecasts has gained attention from central banks over the last years (for instance, the Bank of England produces each quarter a density forecast or 'fan chart' of inflation). As in the case of point forecasts, the researcher may have several models (or experts) generating probabilistic forecasts. It thus seems natural to combine the resulting individual forecasts to generate a single probabilistic forecast.

Recent applications of probabilistic forecasts combination include Geweke and Amisano (2011b) and Jore *et al.* (2010).[18] Clemen and Winkler (2007) and Primo *et al.* (2009) provide comprehensive reviews on how to combine probabilistic forecasts; hence, we only present here a brief overview of the main alternatives available in the literature.

The most common approach to probabilistic forecast combination is the linear opinion pool (Stone, 1961), which is based on a weighted linear combination of the individual probabilistic forecasts. In particular, consider J probabilistic forecasts h periods ahead made by model or expert j $(j = 1, \ldots, J)$ of a variable y at period T, denoted by $p_j(y_{T+h})$. The linear opinion pool would produce a finite mixture resulting from the following combination of the individual density forecasts:

$$p(y_{T+h}) = \sum_{j=1}^{J} w_j p_j(y_{T+h}) \tag{40}$$

where w_j are a set of non-negative weights that sum to unity. These weights are determined according to some optimality criteria, and can receive numerous interpretations (see Genest and McConway (1990)). For instance, Hall and Mitchell (2007) propose to use weights that minimize the Kullback–Leibler information criterion between the combined forecast density and the true (but unknown) density.

There is overwhelming empirical evidence that the linear opinion pool performs better than individual forecasts. However, in the framework of binary outcomes,[19] Ranjan and Gneiting (2010) demonstrates theoretically and empirically that the linear opinion pool is suboptimal because it is uncalibrated[20] (see their Theorem 1 in p. 73). In order to address this caveat, Ranjan and Gneiting (2010) propose a non-linear generalization that nests the traditional, linearly combined probability forecast; more concretely, the authors introduce the so-called beta-transformed linear pool which fits an optimally recalibrated forecast combination, by compositing a beta transform and the traditional linear opinion pool.

Other typical combination approaches are based on multiplicative averaging (sometimes denominated logarithmic opinion pool) or combining log odds ratios by logit regressions (Kamstra and Kennedy, 1998). Another approach is the mixture modelling in Geweke and Amisano (2011a), who construct density forecasts of daily S&P returns by means of a generalized Markov normal mixture model estimated by Bayesian methods.

Finally, inspired by the situation in which the researcher (or decision maker) is confronted with subjective probability distribution of several experts, Genest and Zidek (1986) argues that a Bayesian updating scheme is the most appropriate method to combine them in the typical risk analysis situation. The key intuition of the approach is that the researcher should use Bayes' theorem to update a prior distribution with the information provided by the experts (e.g. Morris, 1974, 1977).

The BMA approach described in Section 2 is also a natural scheme for combining density forecasts (e.g. Garratt *et al.*, 2003). The predictive density can be computed by averaging over the set of candidate models available for forecasting as follows:

$$f(y_{T+h}|y) = \sum_{j=1}^{2^q} f(y_{T+h}|y, M_j) P(M_j|y) \tag{41}$$

where $f(y_{T+h}|y, M_j)$ is the predictive density conditional on model M_j given available data y, and $P(M_j|y)$ is the posterior model probability as described above. In other words, the optimal density forecast from the BMA viewpoint is given by the combination of model-specific forecasts using posterior model probabilities as weights.

The use of the posterior model probabilities as weights might result in in-sample overfitting of the data when the prior contains little information, as it is typically the case in the model averaging framework.

Instead, we can consider measures of out-of-sample performance by substituting the marginal likelihood in the calculation of $P(M_j|y)$ (see equation (4)) by the predictive likelihood, which is $f(y_{T+h}|y, M_j)$ once y_{T+h} has been observed (see Geweke and Whiteman (2006) p. 15–16). In practice, this procedure is implemented by splitting the sample $y_1, \ldots, y_T$ into two parts, one used to convert parameter priors into posterior distributions, and another part for evaluating the out-of-sample forecasting model performance.

In a recent paper, Geweke and Amisano (2011b) illustrate that the performance of optimal pools (i.e. linear opinion pools with weights computed in the spirit of Hall and Mitchell (2007)) is very different from those constructed by means of BMA. In particular, optimal pools perform better than BMA in terms of the log predictive score function.[21] It is worth stressing that large samples are key to superior performance from optimal pools in practice (e.g. Geweke (2010)). Finally, Geweke and Amisano (2011b) show that the assumption of a complete model space made by BMA is at the root of these results; in other words, BMA requires that the true model is among the models available for combination (and thus BMA weights tend to eliminate all models but the true one as sample size increases), whereas optimal prediction pools make no such an assumption.

6. Model Averaging and Endogeneity

The methods described are based on the strict exogeneity assumption of the regressors. This assumption implies that there is no correlation between the variables X and the unobservables (ϵ) affecting the output Y (i.e. $\text{cov}(X, \epsilon) = 0$ in equation (2) above).

However, in many applications this assumption might be implausible. As a result, we may have some variables X say X_1, that are endogenous, and some others, say X_2, that are exogenous (i.e. X_1 variables are correlated with the unobservables given the X_2 variables and thus $\text{cov}(X_1, \epsilon|X_2) \neq 0$). Under these circumstances, obtaining estimates of causal effects requires the availability of an exogenous source of variation on the endogenous variables, that is, a set of valid instruments (Z) which satisfies the conditional IV identifying assumption $\text{cov}(Z, \epsilon|X_2) = 0$.

Given the interest on causal effects over the last decades, how to tackle the issue of endogeneity in the model averaging framework is an important line of open research (e.g. Durlauf *et al.* (2008), Eicher *et al.* (2012b)).

Formally, when we face a situation in which we have endogenous (X_1) and exogenous (X_2) regressors together with a set of valid instruments (Z) in a linear context, the model to be estimated is:

$$\begin{aligned} y &= X_1\beta_1 + X_2\beta_2 + \epsilon \\ X_1 &= Z\pi_1 + X_2\pi_2 + V \end{aligned} \tag{42}$$

where y and X_1 are the $N \times 1$ vector and the $N \times q_1$ matrix of endogenous variables, X_2 is the $N \times q_2$ matrix of exogenous regressors or control variables[22] and Z corresponds to the $N \times q_Z$ matrix of instrumental variables. Moreover, β_1, β_2, π_1 and π_2 represent the $q_1 \times 1$, $q_2 \times 1$, $q_Z \times q_1$ and $q_2 \times q_1$ vectors and matrices of parameters, respectively. Finally, the unobservables in the first equation (i.e. the structural form equation) are captured by the $N \times 1$ vector ϵ, and V is the $N \times q_1$ matrix of errors corresponding to the q_1 remaining equations (usually labelled as reduced form equations).

In this framework, we can define the $Q \times 1$ vector $U_i = (\epsilon_i, V_i')'$ and further assume:

$$U_i \sim N(0, \Sigma) \tag{43}$$

where Σ is a $Q \times Q$ symmetric and positive definite covariance matrix, and $Q = 1 + q_1$. Given this assumption we can construct the likelihood function for such a model under Gaussian errors and estimate the parameters via maximum likelihood (i.e. Limited Information Maximum Likelihood (LIML)),[23] or we can estimate the parameters via method-of-moments estimators (e.g. 2SLS). In both cases we need to have as many instruments as endogenous regressors ($q_Z \geq q_1$) together with the rank condition

$rank(E(Z'X_1)) = q_1$ in order to guarantee identification. In the just-identified case ($q_Z = q_1$), LIML and 2SLS coincide.

Given the IV setting described earlier, two main sources of model uncertainty arise. In particular we might have uncertainty surrounding the selection of endogenous variables X_1 of interest, and uncertainty in the choice of exogenous (or control) variables X_2. As previously stated, how to address the problem of model uncertainty in these settings is an open issue in the model averaging literature. Given the LIML likelihood function and following techniques advanced by Raftery (1995), one natural possibility is the combination of LIML estimates with BIC model weights (see Section 2.4.3).

An important remark here is the importance of considering comparable likelihoods across models. Even in the case of a model not including some elements of X_1 in the structural equation, for the sake of comparability we need to consider the full set of reduced form equations for all the endogenous variables in X_1. This means that we must construct for all the models the same likelihood $f(y, X_1|X_2, Z)$ in order to guarantee comparability across all the models under consideration, that is, the joint likelihood of y and X_1 must be constructed for all the models.

The differences across models emerge in the form of zero restrictions on the parameter vectors β_1, β_2 and π_2 for those variables not included in a particular model (either X_1 for β_1, or X_2 for β_2 and π_2). However, the key point is that the set of q_1 reduced form equations for X_1 must be considered in all the candidate models (i.e. the matrix π_1 is the same in all the models) despite not all the q_1 endogenous variables in X_1 are included in all the models' structural form equations given the existence of model uncertainty in the choice of such variables.

6.1 *Recent Advances in the Literature*

Durlauf *et al.* (2008) represents the first attempt to address the issue of endogenous regressors in a BMA context.[24] More concretely, the authors are concerned with uncertainty surrounding the selection of the endogenous and exogenous variables of interest. Therefore they consider $2^{q_1+q_2}$ candidate models indexed by $j = 1, \ldots, 2^{q_1+q_2}$. The authors propose to use 2SLS model-specific estimates for each single model, and then take the average:

$$E(\theta|y) = \sum_{j=1}^{2^{q_1+q_2}} P\left(M_j|y\right) E(\theta|y, M_j) \approx \sum_{j=1}^{2^{q_1+q_2}} P\left(M_j|y\right) \widehat{\theta}^j_{2SLS} \tag{44}$$

where $\widehat{\theta}^j_{2SLS}$ is the 2SLS estimate for model j, and $\theta = (\beta_1, \beta_2, vec(\pi_1), vec(\pi_2), vech(\Sigma))$ is the $h \times 1$ vector of parameters to be estimated. The weights $P\left(M_j|y\right)$ are inspired in a limited information version of the BIC (i.e. LIBIC) approximation to the integrated likelihood. Finally, regarding the priors on the model space, Durlauf *et al.* (2008) consider the dilution priors.

The formal justification of this approach remains an open issue as stated by the authors. In particular, if an endogenous variable is not included in a particular model, its associated reduced form equations are not included either. As a result, the model posterior probabilities are based on pseudo-likelihoods that might not be fully comparable across models (i.e. the likelihood $f(y, x_1|z)$ is not fully comparable to the likelihood $f(y, x_2|z)$).

More recently, Durlauf *et al.* (2011) consider model averaging across just-identified models. In this case, model-specific 2SLS estimates $\widehat{\theta}^j_{2SLS}$ coincide with model-specific LIML estimates $\widehat{\theta}^j_{LIML}$; hence, likelihood-based BIC weights have formal justification.

In a recent paper, Eicher *et al.* (2012b) 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 (Z). This third source of uncertainty emerges if we have a set of instruments that satisfy all the exclusion restrictions given a set of endogenous variables regardless of the particular model considered.[25] Eicher *et al.* (2012b) propose a 2-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. In both steps the authors propose to use BIC weights.

6.2 *Panel Data and Model Averaging*

Another relevant open line of research is that of model averaging and panel data as an alternative approach to address the issue of endogenous regressors. Omitted variables biases arising from individual-specific and time-invariant unobservable factors can be alleviated by resorting to panel data models with fixed effects. Panel data comprises information on individuals ($i = 1, \ldots, N$) over different time periods ($t = 1, \ldots, T$). Therefore, the correlation between X and ϵ might arise due to the existence of a time-invariant and individual-specific component of ϵ_{it} ($\epsilon_{it} = \eta_i + \vartheta_{it}$) so that:

$$y_{it} = x_{it}\beta + \eta_i + \vartheta_{it} \tag{45}$$

Within the BMA framework, Moral-Benito (2012a) considers such a panel setting with strictly exogenous regressors (uncorrelated with $\vartheta_{i1}, \ldots, \vartheta_{iT}$ but correlated with the permanent component of the error term η_i).

In this framework, the vector x_{it} can also include a lagged dependent variable (y_{it-1}) which is correlated with ϑ_{it-1}. Moral-Benito (2012a) also considers such a dynamic panel model within the BMA approach by combining the likelihood function discussed in Alvarez and Arellano (2003) with the g-UIP priors.

Finally, panel data can also be useful for addressing concerns on reverse causality from the dependent variable to the x regressors. In particular, we can allow for reverse causality (or feedback) by assuming that realizations of the regressors far enough in time are independent of the current shocks. Along these lines, Moral-Benito (2013) constructs a likelihood function for panel data models with unobserved heterogeneity and endogenous regressors that can be combined with BMA techniques. The same panel setting is also considered in Chen *et al.* (2009) who combine panel GMM estimators with BMA. In particular, they interpret the exponentiated GMM objective function as the model-specific pseudo-marginal likelihood, and then use LIBIC weights in the spirit of Durlauf *et al.* (2008).

7. A Brief Overview of the Literature on Model Averaging with Applications to Economics

Until the nineties, the bulk of the literature on model averaging came from two different sources: on the one hand, statistical papers developing the BMA apparatus with little emphasis on economic applications (e.g. Raftery (1995), Volinsky *et al.* (1997) and Fernández *et al.* (2002)), and, on the other hand, papers from the forecasting combination literature.

However, since the beginning of the twenty-first century, new methods together with more powerful computers are inspiring a flurry of BMA activity in different fields of economics. Geisel (1973) and Moulton (1991) represent two exceptions of economic research considering model averaging previous to the 'BMA revolution' in the late nineties. Geisel (1973) compared the prediction ability of macromodels based on posterior model probabilities of consumption equations; on the other hand, Moulton (1991) applied model averaging to 4096 hedonic regressions of quality-adjusted price index numbers for radio services in order to disentangle which characteristics were more important as price determinants.

Brock *et al.* (2003, 2006) highlight the importance of rethinking how to formulate and present policy advice in economics when model uncertainty is present. They embed model uncertainty and policy evaluation in a decision–theoretic framework and consider model averaging techniques to empirically address these issues in the field of monetary policy. In particular they consider 25, 600 different models given by a Taylor rule equation for the interest rate together with different IS and Phillips curve equations determining the output gap and the inflation rate. The different models here come from the inclusion of different lags of interest rates, inflation and output gap in the IS and Phillips curve equations (see also Onatski and Stock (2002) and Onatski and Williams (2003)).

Empirical growth is, without any doubt, the most active field in which model averaging techniques are being applied.[26] The seminal papers on model averaging and growth are Fernández *et al.* (2001b) and Sala-i-Martin *et al.* (2004). Following Raftery (1995), Sala-i-Martin *et al.* (2004) combine OLS estimates with BIC weights, and, Fernández *et al.* (2001b) employ the Benchmark g Priors on the parameter space (see also Masanjala and Papageorgiou (2008) and Crespo-Cuaresma *et al.* (2009)). Also in the growth context, Magnus *et al.* (2010) consider the WALS approach, and Wagner and Hlouskova (2009) apply an FMA estimator based on principal components using four weighting schemes: equal, MMA, AIC and BIC. In the field of growth empirics, we can also find the first attempts to tackle endogeneity concerns within the model averaging framework. See for instance Durlauf *et al.* (2011), Chen *et al.* (2009) and Moral-Benito (2012a) as summarized in Section 6.

Model averaging has also been considered in macroeconomic applications such as forecasting output growth (e.g. Min and Zellner (1993) and Koop and Potter (2004)) or exchange rates (e.g. Wright 2008a)); Garratt *et al.* (2003) also employ BMA to predict inflation and output growth in the UK while Wright (2008b) forecasts US inflation; Garratt *et al.* (2009) study the effect of money on inflation and output in the UK based on a BMA-cointegration approach; Koop *et al.* (1997) investigate the persistence properties of GNP in the USA by averaging inference over ARFIMA and ARMA models.

Also by means of model averaging, Crespo-Cuaresma and Slacik (2009) identify the most important determinants of currency crises in the framework of binary choice models for a panel of countries, and Eicher *et al.* (2012a) study the effect of Preferential Trade Agreements (PTAs) on trade flows using BMA.

With respect to the field of labour/migration, Koske and Wanner (2013) study the drivers of income inequality in OECD countries by means of BMA; Tobias and Li (2004) apply model averaging to estimate Mincer equations; and, Mitchell *et al.* (2011) investigate the determinants of international migration to the UK combining panel data with model averaging techniques.

In finance, Pesaran *et al.* (2009) employ model averaging as a remedy to the risk of inadvertently using false models in portfolio management. Avramov (2002) and Cremers (2002) predict stock returns based on model averaging, and both papers report improved forecasting performance from BMA. Other studies applying model averaging in finance, such as Geweke and Amisano (2011b) and Geweke and Amisano (2011a), are summarized in Section 5.

In other fields of economics, Wan and Zhang (2009) consider FMA estimators to determine the degree to which recreation and tourism development affected a range of socioeconomic indicators (e.g. earnings per job, income per capita) in 311 rural US counties in the 1990s and 2000; Cohen-Cole *et al.* (2009) study the controversial issue of the deterrent effect of capital punishment employing a BMA approach as discussed in Section 4; Galbraith and Hodgson (2009) analyse the determinants of the value of works of art using model averaging. Finally, in health economics, Morales *et al.* (2006) characterize the dose–response relationship between an environmental exposure and adverse health outcomes using model averaging techniques.

7.1 *Software for Model Averaging*

Most of the model averaging methods surveyed in this paper can be implemented in practice using software freely available on the internet. The following webpages can be useful for applied researchers and practitioners interested in model averaging:

1. In the webpage http://bms.zeugner.eu/, Martin Feldkircher and Stefan Zeugner provide a tutorial and a free R package for performing BMA under different priors. This webpage also provides useful links to other BMA resources on the internet.

2. Hoeting *et al.* (1999) summarize several BMA programs, written in S-Plus, available at http://www2.research.att.com/∼volinsky/bma.html.
3. In the webpage http://www.warwick.ac.uk/go/msteel/, Mark Steel provides the data and Fortran codes used in the papers Fernández *et al.* (2002); Fernández *et al.* (2001b) and Ley and Steel (2007).
4. Jan Magnus provides codes, written in Matlab, for implementing the WALS approach (Magnus *et al.* (2010)). He also makes available two STATA commands (`bma` and `wals`) that implement BMA with benchmark priors and WALS (see De Luca and Magnus (2011) for a description of these commands).
5. Finally, Bruce Hansen provides in his webpage several codes (in R, Matlab, Gauss and STATA) implementing both MMA and JMA (see http://www.ssc.wisc.edu/∼bhansen/progs/progs_ma.html).

8. Concluding Remarks

It is common in empirical research to present one baseline specification and several robustness checks in a companion table or even in an appendix. Researchers typically base their conclusions on this baseline specification acting as if the model chosen is the true model. This procedure tends to produce excessively optimistic conclusions due to the underestimation of the uncertainty associated with the whole estimation procedure. This is so because uncertainty surrounding the selection of the empirical model (i.e. model uncertainty) is basically ignored.

This situation represents a challenge to empirical researchers because, as illustrated by Leamer (1983), conclusions from empirical studies may well depend on the selected model (controls variables included). Hence, the results are sensitive to different choices of control explanatory variables. Model averaging approaches estimate the effect of interest under all the possible combinations of controls, and report a weighted average effect. Therefore, model averaging takes into account the uncertainty surrounding the selection of controls (i.e. model uncertainty) in a natural manner.

This paper has presented an overview of existent model averaging techniques and their applications in economics. The usefulness and practical implementation of the methods surveyed in the paper have been illustrated in an empirical application. Using data on murders and executions across states in the USA, I have found that model uncertainty in this setting is large enough to preclude us from disentangling whether capital punishment has a deterrent effect.

How to tackle the issue of endogenous regressors in the model averaging framework is an interesting line of open research. The state-of-the-art of the literature on BMA and endogeneity has been summarized. Allowing for endogenous regressors in the FMA approach could be an interesting topic for future research.

Acknowledgements

I would like to thank Manuel Arellano for provoking my interest in the topic, and Stéphane Bonhomme, Eduardo Ley, Enrique Sentana and two anonymous referees for helpful comments. Research funding from the Spanish Ministry of Science and Innovation, Consolider Grant CSD2006-00016 is gratefully acknowledged.

Notes

[1] In general, ignoring model uncertainty results in narrower confidence bands for the estimate of interest.

[2] Moreover, the focus of this paper is on settings in which the model space is complete (i.e. the ‘true’ model is among the 2^q candidate models under consideration). However, we hold the view that we cannot identify the ‘true’ model from the pool of all candidate models. Draper (1995) provides a

discussion on the existence of a 'true' model and its implications (see also Geweke and Amisano (2011b)).

3. On June 29, 1972, the Supreme Court imposed a moratorium on capital punishment that ended in July 1976.
4. The data are publicly available at http://bpp.wharton.upenn.edu/jwolfers/DeathPenalty.shtml.
5. Despite the use of BMA in applied economics, research was not popularized until the late nineties, model averaging has been present in Bayesian statistics well before.
6. Note that Section 7 is entirely devoted to summarize the literature on model averaging applications to economic research.
7. In order to calculate the posterior model probability in (4) we also need to compute $f(y|M_j)$ that is often called the marginal (or integrated) likelihood.
8. How to choose the particular training sample is a controversial issue (see Berger and Pericchi (1993) and O'Hagan (1995)).
9. Note that the focus regressors may only include a constant term so that we may have the same situation as in the previous section in which all the regressors were focus regressors.
10. For instance, Doppelhofer (2008) considers the EB approach as an alternative to BMA rather than a particular choice of priors within the BMA approach.
11. Eicher *et al.* (2009) compare different prior structures and conclude that the combination of the UIP on the parameter space and the uniform prior on the model space (see the next subsection about priors on model space) outperforms any other possible combination of priors previously considered in the BMA literature in terms of cross-validated predictive performance.
12. To some extent, the dilution priors on model space described in the previous section take into account these interdependencies among redundant regressors. However, you need to elicit the priors before seeing the data; hence you have to assume if the regressors are substitutes or complements *ex ante*.
13. The equivalence of classical inference and Bayesian inference under diffuse priors is well known in the classical normal regression model. For the LIML case, Kleibergen and Zivot (2003) show this equivalence for a particular choice of non-informative priors. Note also that the large sample equivalence is only an approximation.
14. Moreover, as noted by Moral-Benito (2012b), posterior distributions of the parameters can also be obtained within this approach. Analogously to the posterior mean, these posterior distributions are weighted averages of marginal posterior distributions conditional on each individual model. More concretely, these posteriors are mixture normal distributions because model-specific posteriors are normal. This is so because we can make use of the Bernstein-von Mises theorem (see Berger, 1985) which basically states that a Bayesian posterior distribution is well approximated by a normal distribution with mean at the MLE and dispersion matrix equal to the inverse of the Fisher information.
15. Note that the resulting 65, 536 model-specific estimates of β are plotted in Figure 1.
16. Note that these standard errors are computed following (7) in columns (1)–(3) and (27) in columns (4)–(6); therefore, they incorporate not only the uncertainty associated to the parameter estimate conditional on a given model, but also the uncertainty across estimates from different models.
17. For the sake of brevity, the reader is referred to the Timmermann (2006) survey for more details on other point forecasts combination schemes based on asymmetric loss functions, non-parametric methods, time-varying weights or trimming approaches.
18. The combination of probabilistic forecasts is also present in areas such as meteorology (e.g. Raftery *et al.*, 2005).
19. Note that, in the case of a future binary event, such as inflation above a certain threshold versus inflation below that threshold, the predictive pdf is simply the probability for the event to occur.
20. The goal in probability forecasting is to maximize the sharpness (or how far away the forecasts are from the naive forecast, i.e. the closer to the most confident values of 0 or 1, the sharper the forecast)

of the probabilistic forecasts subject to calibration (or how close conditional event frequencies are to the forecast probabilities) (see Gneiting *et al.* (2007)).

21. Note that the optimal prediction pool approach as set out by Geweke and Amisano (2011b) is only optimal from the perspective of the log predictive score (the log score is analogous to model assessments based on log likelihood or marginal likelihood in the conventional frequentist or Bayesian settings, respectively. Also, the analysis is entirely out-of-sample). However, note also that a decision–theoretic approach would not generally deliver this as optimal in the sense of minimizing the decision maker's loss function.
22. We can also refer to the exogenous regressors X_2 as control or conditioning variables in the sense that, in some cases, they must be included in the model in order to guarantee the validity of the instruments even if their effect is not of central interest.
23. Note that the resulting ML estimator can also be interpreted as a pseudo-ML estimator under non-normality.
24. Despite the section is devoted to the connection between model averaging and endogeneity, all advances on this direction are based on the Bayesian spirit of model averaging.
25. In the Durlauf *et al.* (2011) setting, the inclusion of a given endogenous variable in the model implied its own set of valid instruments to be included in the model.
26. Given the limited number of observations (i.e. countries) and the large set of candidate growth determinants, the fragility of these regressions causes a big concern among growth researchers.

References

Alvarez, J. and Arellano, M. (2003) The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica* 71: 1121–1159.

Avramov, D. (2002) Stock return predictability and model uncertainty. *Journal of Financial Economics* 64: 423–258.

Barnard, G. (1963) New methods of quality control. *Journal of the Royal Statistical Society Series A* 126: 255–258.

Bartlett, M. (1957) A comment on D.V. Lindleys statistical paradox. *Biometrika* 44: 533–534.

Bates, J. and Granger, C. (1969) The combination of forecasts. *Operational Research Quarterly* 20: 451–468.

Berger, J. (1985) *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, J. and Pericchi, L. (1993) The intrinsic Bayes factor for model selection. Technical Report 93-43C, Department of Statistics, Purdue University, West Lafayette.

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

Brock, W., Durlauf, S. and West, K. (2006) Model uncertainty and policy evaluation: some theory and empirics. *Journal of Econometrics* 136: 629–664.

Buckland, S., Burnham, K. and Augustin, N. (1997) Model selection: an integral part of inference. *Biometrics* 53: 603–618.

Chen, H., Mirestean, A. and Tsangarides, C. (2009) Limited information Bayesian model averaging for dynamic panels with short time periods. Working Paper WP/09/74, IMF.

Claeskens, G. and Hjort, N. (2003) The focused information criterion. *Journal of the American Statistical Association* 98: 900–916.

Claeskens, G. and Hjort, N. (2008) *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Clemen, R. (1989) combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* 5: 559–583.

Clemen, R. and Winkler, R. (2007) Aggregating probability distributions. In E. Ward, R. Miles and D. von Winterfeldt (eds.), *Advances in Decision Analysis: From Foundations to Applications* (pp. 154–176). Cambridge: Cambridge University Press.

Clyde, M. and George, E. (2000) Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistics Society Series B* 62: 681–698.

Cohen-Cole, E., Durlauf, S., Fagan, J. and Nagin, D. (2009) Model uncertainty and the deterrent effect of capital punishment. *American Law and Economics Review* 11: 335–369.

Cremers, K. (2002) Stock return predictability: a Bayesian model selection perspective. *The Review of Financial Studies* 15: 1223–1249.

Crespo-Cuaresma, J. and Slacik, T. (2009) On the determinants of currency crises: the role of model uncertainty. *Journal of Macroeconomics* 31: 621–632.

Crespo-Cuaresma, J., Doppelhofer, G. and Feldkircher, M. (2009) The determinants of economic growth in European regions. Working Paper Series, CESifo.

Davis, W. (1979) Approximate Bayesian predictive distributions and model selection. *Journal of the American Statistical Association* 74: 312–317.

De Luca, G. and Magnus, J. (2011) Bayesian model averaging and weighted-average least squares: equivariance, stability, and numerical issues. *Stata Journal* 11: 518–544.

Dezhbakhsh, H. and Shepard, J. (2006) The deterrent effect of capital punishment: evidence from a judicial experiment. *Economic Inquiry* 44: 512–535.

Donohue, J. and Wolfers, J. (2006) Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Review* 58: 791–846.

Doppelhofer, G. (2008) Model Averaging. in *The New Palgrave Dictionary in Economics* ed. by L. Blume and S. Durlauf.

Doppelhofer, G. and Weeks, M. (2009) Jointness of growth determinants. *Journal of Applied Econometrics* 24: 209–244.

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

Durlauf, S., Kourtellos, A. and Tan, C. (2008) Are any growth theories robust? *Economic Journal* 118: 329–346.

Durlauf, S., Kourtellos, A. and Tan, C. (2011) Is God in the details? A reexamination of the role of religion in economic growth. *Journal of Applied Econometrics* 27: 1059–1075.

Edgerton, H. and Kolbe, L. (1936) The method of minimum variation for the combination of criteria. *Psychometrika* 1: 183–188.

Ehrlich, I. (1975) The deterrent effect of capital punishment: a question of life and death. *American Economic Review* 65: 397–417.

Eicher, T., Papageorgiou, C. and Raftery, A. (2009) Default Priors and Predictive Performance in Bayesian Model Averaging, with Application to Growth Determinants. *Journal of Applied Econometrics* 26: 30–55.

Eicher, T., Henn, C. and Papageorgiou, C. (2012a) Trade creation and diversion revisited: accounting for model uncertainty and natural trading partner effects. *Journal of Applied Econometrics* 27: 296–321.

Eicher, T., Lenkoski, A. and Raftery, A. (2012b) Bayesian model averaging and endogeneity under model uncertainty: an application to development determinants. *Econometric Reviews*, forthcoming.

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

Fernández, C., Ley, E. and Steel, M. (2001b) Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16: 563–576.

Fernández, C., Ley, E. and Steel, M. (2002) Bayesian modeling of catch in a northwest Atlantic fishery. *Journal of the Royal Statistical Society Series C* 51: 257–280.

de Finetti, B. (1972) *Probability, Induction, and Statistics* New York: Wiley.

Foster, D. and George, E. (1994) The risk inflation criterion for multiple regression. *The Annals of Statistics* 22: 1947–1975.

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

Galbraith, J. and Hodgson, D. (2009) Dimension reduction and model averaging for estimation of artists' age-valuation profiles. Working Paper, CIRANO.

Garratt, A., Lee, K., Pesaran, H. and Shin, Y. (2003) Forecast uncertainties in macroeconomic modeling: an application to the U.K. economy. *Journal of the American Statistical Association* 98: 829–838.

Garratt, A., Koop, G., Mise, E. and Vahey, S. (2009) Real-time prediction with U.K. monetary aggregates in the presence of model uncertainty. *Journal of Business & Economic Statistics* 27: 480–491.

Geisel, M. (1973) Bayesian comparisons of simple macroeconomic models. *Journal of Money, Credit and Banking* 5: 751–772.

Geisser, S. (1965) A Bayes approach for combining correlated estimates. *Journal of the American Statistical Association* 60: 602–607.

Geisser, S. and Eddy, W. (1979) A predictive approach to model selection. *Journal of the American Statistical Association* 74: 153–160.

Genest, C. and McConway, K. (1990) Allocating the weights in the linear opinion pool. *Journal of Forecasting* 9: 53–73.

Genest, C. and Zidek, J. (1986) Combining probability distributions. A Critique and annotated bibliography. *Statistical Science* 1: 114–148.

George, E. (1999) Discussion of 'Bayesian Model Averaging and Model Search Strategies' by M. Clyde. In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics 6*. Oxford: Oxford University Press.

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

Geweke, J. (2010) *Complete and Incomplete Econometric Models*. Princeton, NJ: Princeton University Press.

Geweke, J. and Amisano, G. (2011a) Hierarchical Markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics* 26: 1–29.

Geweke, J. and Amisano, G. (2011b) Optimal prediction pools. *Journal of Econometrics* 164: 130–141.

Geweke, J. and Whiteman, C. (2006) Bayesian forecasting. In G. Elliot, C. Granger and A. Timmerman, A (eds.), *Handbook of Economic Forecasting*, Vol. 1, (pp. 3–80). Elsevier: Amsterdam.

Gneiting, T. (2008) Editorial: probabilistic forecasting. *Journal of the Royal Statistical Society Series A* 171: 319–321.

Gneiting, T., Balabdaoui, F. and Raftery, A. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B* 69: 243–268.

Granger, C. and Ramanathan, R. (1984) Improved methods of combining forecasts. *Journal of Forecasting* 3: 197–204.

Granger, C., White, H. and Kamstra, M. (1989) Interval forecasting: an analysis based upon ARCH-quantile estimators. *Journal of Econometrics* 40: 87–96.

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

Halperin, M. (1961) Almost linearly-optimum combination of unbiased estimates. *Journal of the American Statistical Association* 56: 36–43.

Hansen, B. (2007) Least squares model averaging. *Econometrica* 75: 1175–1189.

Hansen, B. (2008) Least squares forecast averaging. *Journal of Econometrics* 146: 342–350.

Hansen, B. and Racine, J. (2012) Jackknife model averaging. *Journal of Econometrics* 167: 38–46.

Hansen, M. and Yu, B. (2001) Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96: 746–774.

Hjort, N. and Claeskens, G. (2003) Frequentist model average estimators. *Journal of the American Statistical Association* 98: 879–899.

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

Horst, P. (1938) Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika* 1: 53–60.

Jore, A., Mitchell, J. and Vahey, S. (2010) Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics* 25: 621–634.

Kamstra, M. and Kennedy, P. (1998) Combining qualitative forecasts using logit. *International Journal of Forecasting* 14: 83–93.

Kass, R. and Wasserman, L. (1995) A reference Bayesian test for nested hypothesis with large samples. *Journal of the American Statistical Association* 90: 928–934.

Katz, L., Levitt, S. and Shustorovich, E. (2003) Prison conditions, capital punishment, and deterrence. *American Law and Economics Review* 5: 318–343.

Kleibergen, F. and Zivot, E. (2003) Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics* 114: 29–72.

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

Koop, G. and Potter, S. (2004) Forecasting in dynamic factor models using Bayesian model averaging. *The Econometrics Journal* 7: 550–565.

Koop, G., Ley, E., Osiewalski, J. and Steel, M. (1997) Bayesian analysis of long memory and persistency using ARFIMA models. *Journal of Econometrics* 76: 149–169.

Koske, I. and Wanner, I. (2013) The drivers of labour income inequality: an analysis based on Bayesian model averaging. *Applied Economics Letters* 20: 123–126.

Kuersteiner, G. and Okui, R. (2010) Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78: 697–718.

Laplace, P.S. (1818) *Deuxime Supplément a la Théorie Analytique des Probabilités*. Paris: Courcier.

Leamer, E. (1978) *Specification Searches*. New York: John Wiley & Sons.

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

Lempers, F. (1971) *Posterior Probabilities of Alternative Linear Models*. Rotterdam: University Press.

Ley, E. and Steel, M. (2007) Jointness in Bayesian variable selection with applications to growth regressions. *Journal of Macroeconomics* 29: 476–493.

Ley, E. and Steel, M. (2009) On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* 24: 651–674.

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

Madigan, D. and Raftery, A. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89: 1535–1546.

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

Magnus, J., Powell, O. and Prüfer, P. (2010) A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* 154: 139–153.

Masanjala, W. and Papageorgiou, C. (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: 671–682.

Min, C. and Zellner, A. (1993) Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics* 56: 89–118.

Mitchell, J., Pain, N. and Riley, R. (2011) The drivers of international migration to the UK: a panel-based Bayesian model averaging approach. *Economic Journal* 121: 1398–1444.

Moral-Benito, E. (2012a) Determinants of economic growth: a Bayesian panel data approach. *The Review of Economics and Statistics* 94: 566–579.

Moral-Benito, E. (2012b) Growth empirics in panel data under model uncertainty and weak exogeneity. Working Paper 1243, Banco de España.

Moral-Benito, E. (2013) Likelihood-based estimation of dynamic panels with predetermined regressors. *Journal of Business & Economic Statistics* 31: 451–472.

Morales, K., Ibrahim, J., Chen, C. and Ryan, L. (2006) Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. *Journal of the American Statistical Association* 101: 9–17.

Morris, P. (1974) Decision analysis expert use. *Management Science* 20: 1233–1241.

Morris, P. (1977) Combining expert judgments: a Bayesian approach. *Management Science* 23: 679–693.

Moulton, B. (1991) A Bayesian approach to model selection and estimation with application to price indexes. *Journal of Econometrics* 49: 169–193.

O'Hagan, A. (1995) Fractional Bayes factors for model comparison. *Journal of the Royal Statistics Society Series B* 57: 99–138.

Onatski, A. and Stock, J. (2002) Robust monetary policy under model uncertainty in a small model of the US economy. *Macroeconomic Dynamics* 6: 85–110.

Onatski, A. and Williams, N. (2003) Modeling model uncertainty. *Journal of European Economic Association* 1: 1087–1122.

Pesaran, H. and Timmermann, A. (2005) Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of Econometrics* 129: 183–217.

Pesaran, H. and Timmermann, A. (2007) Selection of estimation window in the presence of breaks. *Journal of Econometrics* 137: 134–161.

Pesaran, H., Schleicher, C. and Zaffaroni, P. (2009) Model averaging in risk management with an application to futures markets. *Journal of Empirical Finance* 16: 280–305.

Primo, C., Ferro, C., Jolliffe, I. and Stephenson, D. (2009) Combination and calibration methods for probabilistic forecasts of binary events. *Monthly Weather Review* 137: 1142–1149.

Raftery, A. (1995) Bayesian model selection in social research. *Sociological Methodology* 25: 111–163.

Raftery, A., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* 133: 1155–1174.

Ranjan, R. and Gneiting, T. (2010) Combining probability forecasts. *Journal of the Royal Statistical Society Series B* 72: 71–91.

Roberts, H. (1965) Probabilistic prediction. *Journal of the American Statistical Association* 60: 50–62.

Sala-i-Martin, X., Doppelhofer, G. and Miller, R. (2004) Determinants of long-term growth: a Bayesian Averaging of Classical Estimates (BACE) approach. *American Economic Review* 94: 813–835.

Smith, J. and Wallis, K. (2009) A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics* 71: 331–355.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian Variable selection. *Journal of Econometrics* 75: 317–343.

Spiegelhalter, D. and Smith, A. (1982) Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistics Society Series B* 44: 377–387.

Stigler, S. (1973) Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika* 60: 439–445.

Stone, M. (1961) The linear pool. *Annals of Mathematical Statistics* 32: 1339–1342.

Timmermann, A. (2006) Forecast combinations. In G. Elliott, C. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting* (pp. 135–196). Amsterdam: North-Holland.

Tobias, J. and Li, M. (2004) Returns to schooling and Bayesian model averaging: a union of two literatures. *Journal of Economic Surveys* 18: 153–180.

Volinsky, C., Madigan, D., Raftery, A. and Kronmal, R. (1997) Bayesian model averaging in proportional hazard models: predicting the risk of a stroke. *Applied Statistics* 46: 443–448.

Wagner, M. and Hlouskova, J. (2009) Growth regressions, principal components and frequentist model averaging. Working Paper, Institute for Advanced Studies, Vienna.

Wan, A. and Zhang, X. (2009) On the use of model averaging in tourism research. *Annals of Tourism Research* 36: 525–532.

Wright, J. (2008a) Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics* 146: 329–341.

Wright, J. (2008b) Forecasting US inflation by Bayesian Model Averaging. *Journal of Forecasting* 28: 131–144.

Zellner, A. (1986) On Assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (389–399). Amsterdam: North-Holland/Elsevier.