MPRA
Munich Personal RePEc Archive

Model averaging in economics

Moral-Benito, Enrique

CEMFI

October 2010

Online at https://mpra.ub.uni-muenchen.de/26047/
MPRA Paper No. 26047, posted 21 Oct 2010 16:09 UTC

Model Averaging in Economics*

Enrique Moral-Benito†

CEMFI

October 2010

ABSTRACT

Fragility of regression analysis to arbitrary assumptions and decisions about choice of control variables is an important concern for applied econometricians (e.g. Leamer (1983)). Sensitivity analysis in the form of model averaging represents an (agnostic) approach that formally addresses this problem of model uncertainty. This paper presents an overview of model averaging methods with emphasis on recent developments in the combination of model averaging with IV and panel data settings.

JEL Classification: C5.

Keywords: Model Averaging, Model Uncertainty.

*I would like to thank Manuel Arellano for provoking my interest in the topic, and Enrique Sentana and Stéphane Bonhomme for helpful comments.

†E-mail address: enrique.moral@gmail.com

1 Introduction

In the early Eighties, Leamer (1983) pointed out the fragility of regression analysis to arbitrary decisions about choice of control variables. The selection of control variables in empirical research is an important issue which usually hampers consensus on the particular empirical model to be estimated. The empirical growth literature is probably the best example. In the growth regressions industry, the main area of effort has been the selection of appropriate variables to include in linear growth regressions, resulting in a total of more than 140 variables proposed as growth determinants. Parameter estimates emerging from these regressions are highly fragile to the inclusion of different sets of regressors.

Extreme Bound Analysis (EBA) was first proposed (Leamer (1983), Leamer and Leonard (1983)) as a tool for quantifying the sensitivity of regression estimates. Suppose one is interested in measuring the effect of the variable X on the variable Y. Extreme Bounds Analysis consist of the following steps: first, we estimate a fairly general model in which you regress Y on X and a set of other (control) variables; second, we estimate several simplified versions of the general model (for example by excluding one or more explanatory variables); finally we analyze all the different estimated coefficients on X. Extreme Bounds Analysis is concerned with the largest and smallest values of these estimates. Suppose the estimated coefficient varies greatly over the range of estimated models. Inference concerning the coefficient is then said to be be fragile or unreliable, since the coefficient estimate obtained appears to be sensitive to the precise specification of the model used. Some authors have criticized this approach on the grounds of its ad-hoc nature and because it merely presents in a different format the same information as does conventional regression analysis (e.g. Angrist and Pischke (2010), McAleer et al. (1985)).

Imagine a situation in which there are many different candidate models for estimating the effect of X on Y. Facing this challenge, one can estimate a single model and then make inference based on that selected model ignoring the uncertainty surrounding the model selection process (i.e. the model selection approach). Following this approach we are implicitly assuming not only that there exists a true model, but also that this model is included among the candidate models considered by the researcher. 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). With this approach we do not need

to assume that a true model exists. Moreover, after computing the associated standard errors, one can make inference based on the whole universe of candidate models. By doing so, we would be considering 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. This approach would lead us to more reliable, or at least more honest, conclusions regarding the significance of the estimated effect of X on Y. This paper presents an overview of the literature on model averaging techniques applied to empirical research in economics.

Empirical growth is probably the field in economics in which concerns about fragility of regression estimates have caused the most active literature applying sensitivity analysis techniques. In this context, Levine and Renelt (1992) is possibly the most influential paper considering extreme bound analysis, and Sala-i-Martin et al. (2004) is certainly the paper that popularized the use of model averaging as a tool for applied researchers in the empirical growth regressions literature.

As in any other subfield of econometrics, in this literature there are Bayesian and Frequentist approaches to model averaging. The main differences between Frequentist Model Averaging (FMA) and Bayesian Model Averaging (BMA) arise from the way in which inference is made, and from the selection of model weights. 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 last decade. In this paper I review the state of the art in both approaches providing a discussion of their advantages and drawbacks. Moreover, I argue that the two perspectives are complementary and they can be reinforced to each other so that dogmatic recommendations are not in the agenda of this paper.

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 (IV) models is an interesting line of open research. Panel data represent an alternative to IV for estimating causal effects in situations with endogenous regressors, so extending the model averaging apparatus to panel data models is also a relevant research topic. The first steps on this direction have been taken during the last lustrum (e.g. Durlauf et al. (2008), Moral-Benito (2010a)). In this paper I summarize the recent developments on model averaging with endogenous regressors.

The remainder of the paper is organized as follows. Section 2 intuitively presents the basic concepts of model averaging techniques. The Bayesian approach to model averaging

is formally presented in Section 3, and the Frequentist alternative in Section 4. Section 5 describes recent model averaging approaches to settings with endogenous regressors, and in Section 6 I present some examples of model averaging applications in economics. Finally, Section 7 concludes.

2 Model Averaging as an (Agnostic) Alternative

It is common in empirical papers to have one baseline specification in which the central conclusions are based on, and several robustness checks in a companion table or even in an appendix. The main objective of this approach is to convince the reader that the results in the paper are robust to different assumptions. However, it is still unclear for the readers how hard the authors had to work to find their favorite outcomes. This situation arises because theory does in general not offer enough guidance in the selection of the appropriate empirical model, and then the empirical researcher faces a problem of model uncertainty. Therefore, different researchers might propose and use different empirical models that are compatible with each other for analyzing the same question. In doing so, they first select a single model (for instance based on a given information criterion or a personal view of the question of interest), and second they make inference under the selected model as if this model had been given in advance. In reality, inference based on the final model might give an excessively optimistic answer due to the under-estimation of the uncertainty associated with the whole estimation procedure. Model averaging represents an alternative to this process.[1] Moreover, I consider model averaging as an agnostic approach in the sense that a researcher employing model averaging techniques is unwilling to commit to an opinion about the best single model.

Frequentist Model Averaging (FMA) and Bayesian Model Averaging (BMA) are two different approaches to model averaging in the literature. Despite their similarities in the spirit and the objectives, both techniques differ in the approach to inference and to the selection of model weights. Sections 3 and 4 present an overview of the developments on BMA and FMA over the last years, and how both approaches conduct inference and compute model weights.

[1]Extreme Bound Analysis is also an option to be considered as commented in the Introduction. However, once we accept there are different competing models compatible with each other, model averaging provides a more systematic and rigorous approach to the problem.

2.1 Historical Perspective

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 analyzed the joint distribution of the two, and proposed a combining formula that resulted in a better estimator than either.[2]

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. Halperin (1961) provided a minimum-squared-error combination of estimates, and Geisser (1965) appears to be the earliest Bayesian approach to combining estimates. In the forecasting literature, a flood of papers about combining different forecasts was generated in the 1970s since the seminal papers by Barnard (1963) and Bates and Granger (1969). Timmermann (2006) provides a good overview of this literature.

Despite the basic paradigm for Bayesian Model Averaging (BMA) was introduced by Leamer (1978), the approach was basically ignored until the 1990s and 2000s when there has been an enormous literature on the use of BMA. This is so because more powerful computers and theoretical developments such as Markov chain Monte Carlo Model Composition (MC^3) allow researchers to overcome the troubles related to implementing BMA. Raftery (1995) and Fernández et al. (2001b) are good examples of recent BMA applications in economics. The state of research in the field during the nineties was summarized in Hoeting et al. (1999).

The forecasting combination articles in the 1970s can be considered the predecessors of the current Frequentist Model Averaging (FMA) literature. In contrast to BMA, 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.

[2]See Stigler (1973) for a brief presentation of Laplace's work.

2.2 Basic Concepts

Empirical research in economics is in general plagued by model uncertainty problems. This means that it is very unlikely that only one model needs to be considered. Imagine a researcher who is trying to estimate the effect of a particular policy on a particular outcome. It is a common situation to have more than one possible model to analyze such effect. Let us suppose that the researcher has q possible models in mind, indexed by $h = 1, ..., q$. This implies that there are q different estimates of the effect of interest depending on the model considered, say $\left\{\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_q\right\}$.

In such a situation, the most common approach is to select a single model from the q existing candidates. There is a huge literature on model selection, i.e. the task of selecting a statistical model from a set of potential models given data. A good overview of this literature can be found in Claeskens and Hjort (2008). After the model selection step, both the inference and the conclusions of the analysis are typically based on this single model. For instance, let us think that the selected model is $h = 3$, and therefore the presented result is $\hat{\beta}_3$. On the other hand, it is also very common to present some extra estimates as a robustness check. It is easy to imagine a final table with several columns presenting some of the estimates corresponding to other possible models, for example $\hat{\beta}_{10}$, $\hat{\beta}_{23}$, $\hat{\beta}_{34}$, and $\hat{\beta}_{50}$. If all the presented estimates are close enough, the researcher concludes that her result is robust and the paper is then ready. However, as previously mentioned, it will be still unclear for the readers how hard the researcher had to work to select and defend the selected model $h = 3$, and more importantly, to find the similar estimates $\hat{\beta}_{10}$, $\hat{\beta}_{23}$, $\hat{\beta}_{34}$, and $\hat{\beta}_{50}$ among the q candidates, bearing in mind that we could easily have billions of candidate models.

Model averaging represents an agnostic alternative to this approach. The key idea of model averaging is to consider and estimate all the q candidate models, and then report a weighted average as the estimate of the effect of interest. Therefore, model averaging is an agnostic approach in the sense that a researcher relying on this approach holds the view that the true single model is unknown and probably unknowable. Then, the best she can do is to consider all the possible alternatives instead of selecting one probably incorrect option. The model averaging estimate ($\hat{\beta}_{MA}$) can then be written as:

$$\hat{\beta}_{MA} = \sum_{h=1}^{q} \omega_h \hat{\beta}_h \tag{1}$$

where ω_h represents the weight associated to model h. In subsequent sections we will analyze the different alternatives to choose and estimate the weights ($\hat{\omega}_h$), to obtain the

model-specific estimates $\hat{\beta}_h$, and how can we make inference based on model averaging in both its Bayesian and Frequentist versions.

3 Bayesian Model Averaging

3.1 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:

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

where y and ϵ are $N \times 1$ vectors of the dependent variable and the random shocks respectively. X is a $N \times q$ 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 $\beta = (\beta_1, \beta_2, ..., \beta_q)'$ to be zeros, there are a total of 2^q candidate models to be estimated —indexed by M_j for $j = 1, ..., 2^q$— which all seek to explain y —the data—. For instance, setting β_1 to be zero implies that we are not including the first regressor (i.e. the first column of X, being $X = (X_1, X_2, ..., X_q)$) in the model. 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:

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

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.

On the other hand, Bayesian inference suggests that the posterior model probability can be used to assess the degree of support for M_j. Therefore, posterior model probabilities will be used as model weights in BMA. Given the 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. This means that Bayesian Model Averaging (BMA) involves two different prior beliefs, one on the parameter space ($g(\beta^j|M_j)$) and another one on the model space ($P(M_j)$).

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, and is calculated using (3) and a few simple manipulations. In particular, if we integrate both sides of (3) with respect to β^j, use the fact that $\int g(\beta^j|y, M_j)\, d\beta^j = 1$ (since probability density functions integrate to one), and rearrange, we obtain:

$$f(y|M_j) = \int f\left(y|\beta^j, M_j\right) g\left(\beta^j|M_j\right) d\beta^j. \tag{5}$$

The quantity $f(y|M_j)$ given by equation (5) is the marginal probability of the data, because it is obtained by integrating the joint density of (y, β^j) given y over β^j. The ratio of integrated likelihoods of two different models is the Bayes Factor and it is closely related to the likelihood ratio statistic, in which the parameters β^j are eliminated by maximization rather than by integration.

Following Leamer (1978) we can consider β a function of β^j for each $j = 1, ..., 2^q$ (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:

$$g(\beta|y) = \sum\nolimits_{j=1}^{2^q} P(M_j|y)\, g(\beta|y, M_j) \tag{6}$$

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 $P(M_j|y)$.

Given the Bayesian framework based on parameter distributions, when applying BMA according to equation (6) both estimation and inference come naturally together from the posterior distribution that provides inference about β that takes full account of model uncertainty.

Despite the Bayesian spirit of the approach, one might also be interested in point estimates and their associated variances. If this is so, one common procedure is to take expectations across (6):

$$E(\beta|y) = \sum\nolimits_{j=1}^{2^q} P(M_j|y)\, E(\beta|y, M_j) \tag{7}$$

with associated posterior variance:

$$\begin{aligned} V(\beta|y) &= \sum\nolimits_{j=1}^{2^q} P(M_j|y)\, V(\beta|y, M_j) + \\ &+ \sum\nolimits_{j=1}^{2^q} P(M_j|y) \left(E(\beta|y, M_j) - E(\beta|y)\right)^2 \end{aligned} \tag{8}$$

The posterior variance in (8) incorporates not only the weighted average of the estimated variances of the individual models but also the weighted variance in estimates of the β's across different models. This means that even if we have highly precise estimates in all the models, we might end up with considerable uncertainty about the parameter if those estimates are very different across specifications.

As a by-product of the BMA approach, we can also compute the posterior probability that a particular variable h is included in the regression. In other words, variables with high posterior probabilities of being included are considered as robustly related to the dependent variable of interest. This object is called the *posterior inclusion probability* for variable h, and it is calculated as the sum of the posterior model probabilities for all of the models including that variable:

$$\text{posterior inclusion probability} = P\left(\beta_h \neq 0|y\right) = \sum\nolimits_{\beta_h \neq 0} P\left(M_j|y\right) \tag{9}$$

Implementing Bayesian Model Averaging can be difficult because of two reasons: (i) two types of priors (on parameters and on models) need to be elicited and this can be a complicated task. (ii) the number of models under consideration —2^q— is often huge so 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 problems.

3.2 Priors on the Parameter Space

Prior density choice for Bayesian Model Averaging remains an open area of research. In the context of BMA, improper priors for model-specific parameters cannot be used because they are determined only up to an multiplicative arbitrary constant. Despite these constants cancel in the posterior distribution of the model-specific parameters when doing inference for a given model, they remain in marginal likelihoods leading to indeterminate model probabilities and Bayes factors. To avoid this situation, proper priors for β under each model are usually required. Some of the most popular alternatives considered in the literature are summarized below.

3.2.1 Zellner's g Priors

Given the normal regression framework, the bulk of the BMA literature favors 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:

$$\beta^j|\sigma^2, M_j, g \sim N(0, \sigma^2 g(X_j'X_j)^{-1}) \tag{10}$$

Moreover, the variance parameter σ is common to all the models under consideration, so an improper prior is not problematic, and the most common approach is the uninformative prior proposed by Fernández et al. (2001a): $p(\sigma) \propto \sigma^{-1}$.[3]

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.

Though there are many different options for choosing g (see for example Fernández et al. (2001a)), the three most popular alternatives are:

- 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 Bayesian Information Criterion (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 for example Raftery (1995) or Sala-i-Martin et al. (2004)).

- Risk Inflation Criterion (g-RIC): recommended by Foster and George (1994), it implies setting $g = q^2$.

- 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 $g = max(N, q^2)$.

Eicher et al. (2009c) compare different prior structures and conclude that the combination of the Unit Information Prior 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.

[3]We can also include a constant term (α) in all the models with prior $p(\alpha) \propto 1$.

3.2.2 Laplace Priors

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

$$\begin{aligned} y &= X_1\gamma + X_2\delta + \epsilon \\ \epsilon &\sim N(0, \sigma^2 I_N) \end{aligned} \tag{11}$$

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).[4] We can reparametrize the model in (11) 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 ith element, η_i $(i = 1, ..., q_2)$ is the population t-ratio on δ_i, the ith 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.

3.3 Priors on the Model Space

In order to implement any of the BMA strategies described above, 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 below).

[4]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.

3.3.1 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:

$$\Xi \sim Bin\,(q, \xi) \tag{12}$$

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:

$$P(M_j) = \xi^{q_j}(1-\xi)^{q-q_j} \tag{13}$$

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 (13) reduces to:

$$P(M_j) = 2^{-q} \tag{14}$$

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 ($E(\Xi)$). For instance, the uniform prior just described implies $E(\Xi) = q/2$. The choice of one of the hyperparameters ξ or $E(\Xi)$ automatically produces a value for the other, and it leads to larger or smaller penalizations to big models.

3.3.2 Binomial-Beta Priors

Ley and Steel (2009b) 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:

$$\Xi \sim Bin\,(q, \xi) \tag{15}$$

$$\xi \sim Be\,(a, b) \tag{16}$$

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:

$$E(\Xi) = \frac{a}{a+b} q \tag{17}$$

The model size distribution generated in this way is the so-called Binomial-Beta distribution. Ley and Steel (2009b) propose to fix $a = 1$ and $b = (q - E(\Xi))/E(\Xi)$ through equation (17), 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.

3.3.3 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 neighborhoods of models. Moreover, this kind of dilution prior assigns uniform probability to neighborhoods rather than models, and solves the dilution problem. Consider a given theory (or neighborhood 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:

$$P(C_T) = |R_{C_T}| \prod_{h=1}^{q_T} \xi^{\pi_h} (1-\xi)^{1-\pi_h} \tag{18}$$

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 neighborhood C_T and uniform probability to all the different neighborhoods.

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 neighborhoods) which is usually not within reach.

3.4 Further Topics in BMA

3.4.1 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 have been many algorithms developed 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. Madigan and Raftery (1994) provide a detailed description of the method.

Another commonly-used alternative, initially developed in Madigan and York (1995) is Markov Chain Monte Carlo Model Composition (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 very intuitive and it is not complicated. In the Appendix you can find a detailed description of how does MC^3 work in practice.

3.4.2 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 (7)

as follows:[5]

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

where $\widehat{\beta}_{ML}^{j}$ 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 4.3.1), and given the use of maximum likelihood estimates, I argue that this commonly-used approach to BMA (e.g. Raftery (1995), Sala-i-Martin et al. (2004), Moral-Benito (2010a)) can be considered 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 (2010a) generalized the use of this approach to panel data models in the denominated Bayesian Averaging of Maximum Likelihood Estimates (BAMLE).

Moreover, as noted by Moral-Benito (2010b), posterior distributions of the parameters can also be obtained with 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[6] (also known as the Bayesian CLT) 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.

[5]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.

[6]Berger (1985) provides an in-depth analysis and an excellent illustration.

3.4.3 Jointness

The main focus of Bayesian Model Averaging is the identification of the robust determinants of a given outcome when model uncertainty is present. However, another relevant issue which arises in this framework is to identify whether different sets of regressors are substitutes or complements in the determination of the outcome. For example, in an extreme case, the effect of a particular regressor on the outcome variable might appear or disappear by simply including a specific covariate in the regression. Accounting for these interdependencies among the regressors delivers more parsimonious models with minimally reduced explanatory power. 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 (in the dilution prior setting, a neighborhood is a set of complement covariates). Although in some circumstances you might have an idea, this is not always the case.

In the framework of growth regressions, Ley and Steel (2007) and Doppelhofer and Weeks (2009a) 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.

According to Ley and Steel (2007), any jointness measure should satisfy four criteria: (i) interpretability: any jointness measure should have either a formal statistical or a clear intuitive meaning in terms of jointness; (ii) calibration: values of the jointness measure should be calibrated against some clearly defined scale, derived from either formal statistical or intuitive arguments; (iii) extreme jointness: the situation where two variables always appear together should lead to the jointness measure reaching its value reflecting maximum jointness; and (iv) definition: the jointness measure should always be defined whenever at least one of the variables considered is included with positive probability. Based on these criteria, they propose two alternative measures:

$$J^*_{LS} = \frac{P(i \cap j)}{P(i) + P(j) - P(i \cap j)} \in [0, 1] \tag{21}$$

$$J_{LS} = \frac{P(i \cap j)}{P(i) + P(j) - 2P(i \cap j)} \in [0, \infty) \tag{22}$$

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.

Alternatively, Doppelhofer and Weeks (2009a) propose the following jointness statistic:

$$J_{DW} = \ln \frac{P(i \cap j)P(\widetilde{i} \cap \widetilde{j})}{P(i \cap \widetilde{j})P(\widetilde{i} \cap j)} \tag{23}$$

where $P(i \cap j)$ is the same as before, $P(\widetilde{i} \cap \widetilde{j})$ represents the sum of all the posterior model probabilities of those models in which neither X_i nor X_j are included, and the other two elements are defined accordingly. For more details on the advantages and drawbacks of both approaches see the comments by Strachan (2009) and Ley and Steel (2009a), and the rejoinder of Doppelhofer and Weeks (2009b).

4 Frequentist Model Averaging

4.1 Definition of FMA Estimators

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

$$y = \beta X_1 + X_2 \gamma + U \tag{24}$$

where y, X_1, and U are $N \times 1$ vectors of the dependent variable, the treatment variable of interest and the random shocks respectively. X_2 is a $N \times q$ 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_1 and X_2 for illustration purposes, FMA can easily handle situations in which we cannot make such a distinction. Finally, N is the number of observations in the sample.

If we set some components of $\gamma = (\gamma_1, \gamma_2, ..., \gamma_q)'$ 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 $M \in \{M_1, M_2, ..., M_{2^q}\}$. 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:

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

We can also rewrite the above estimator as

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

where:

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

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 given by:

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

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

4.2 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:[7]

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

where $\Lambda = \sum_{j=1}^{2^q} \omega_{M_j} \Lambda_j$.

However, inference based on this limiting distribution Λ will still ignore the uncertainty involved in the model selection step. Therefore, confidence intervals constructed from $\hat{\beta}_{FMA}$ and the variance of Λ in the usual way, will produce too optimistic inference and will 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

[7] More details about the derivation of this distribution can be found in the Appendix.

standard error of $\hat{\beta}_{FMA}$ is:

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

where $\hat{\tau}_j^2$ estimates the variance of Λ_j, and $\hat{b}_j = \hat{\beta}_{M_j} - \hat{\beta}_{FMA}$ captures the extra uncertainty associated with the variation of estimates across different models. This formula implies an estimated variance for the FMA estimator that closely resembles its Bayesian counterpart in equation (8) (i.e. the posterior variance for the BMA estimator). Note that we still have to replace the fixed weights in equations (25) and (27) by their estimates in order to apply FMA.

4.3 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.

4.3.1 Weight Choice Based on Information Criteria

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

$$I_j = -2\log(L_j) + \varphi_j$$

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).

Buckland et al. (1997) propose to use the following model weights:

$$\omega_{M_j} = \frac{exp(-I_j/2)}{\sum_{h=1}^{2^q} exp(-I_h/2)} \tag{28}$$

The penalty term $\varphi_j = 2q_j$ corresponds to the Akaike Information Criterion (AIC), being q_j the number of parameters in model j. Therefore Akaike weights are one common alternative. Another possible choice is $\varphi_j = q_j \ln(N)$ that corresponds to the Bayesian

Information Criterion (BIC). Given the use of BIC is also justified from a Bayesian viewpoint, this illustrates one clear similarity between BMA and FMA, that provide in fact the same point estimates under some particular circumstances.

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.

4.3.2 Weight Choice Based on Mallow's Criterion

Hansen (2007) proposes to select the model weights in least squares model averaging by minimizing the Mallow's criterion. Despite this criterion is similar to the Akaike information criterion in the model selection spirit, the approach to calculate the weights in Hansen (2007) in the model averaging setting is different.

Hansen (2007) considers the following homoskedastic linear regression:

$$\begin{aligned} y_i &= \sum_{j=1}^{\infty} \theta_j x_{ij} + e_i \\ E(e_i|x_i) &= 0 \\ E(e_i^2|x_i) &= \sigma^2 \end{aligned} \tag{29}$$

where $x_i = (x_{i1}, x_{i2}, ...)$.

Now consider the sequence of candidate models $j = 1, 2, ...$ seeking to approximate (29). The j-th model uses the first ϕ_j elements of x_i with $0 < \phi_1 < \phi_2 < ....$ Given the above, the j-th candidate model is:

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

with corresponding approximating error $\sum_{j=\phi_j+1}^{\infty} \theta_j x_{ij}$. Let us rewrite (30) in matrix form:

$$Y = X_j \Theta_j + e \tag{31}$$

where Y and e are $N \times 1$ vectors, X_j is a $N \times \phi_j$ matrix, and Θ_j is a $\phi_j \times 1$ vector of parameters. Let $J = J(N) \leq N$ be the candidate model with the largest number of

regressors, and $\lambda = (\lambda_1, ..., \lambda_J)'$ a weight vector in the unit simplex in $\mathbb{R}^J$:

$$\mathbf{H}_N = \left\{ \lambda \in [0,1]^J : \sum_{j=1}^{J} \lambda_j = 1 \right\}$$

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

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

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:

$$\hat{\lambda} = \underset{\lambda \in \mathbf{H}_N}{\text{argmin}} \; C_N(\lambda)$$

where:

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

with $\Phi = (\phi_1, ..., \phi_J)'$.

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 $\mathbf{H}_N$ (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.

4.3.3 Weight Choice Based on Cross-Validation Criterion

In a recent paper, Hansen and Racine (2009) 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 labeled 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 this two points, the setup is the same as for

the MMA estimator in (29). Let us further define:

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

so that the jackknife version of the model averaging estimator of μ is:

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

where $\hat{P}_j = \hat{D}_j(P_j - I_N) + I_N$, $P_j = X_j(X_j'X_j)^{-1}X_j'$ is the projection matrix under the j-th candidate model, $\hat{D}_j$ is the $N \times N$ 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:

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

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

$$\hat{\lambda}^* = \underset{\lambda \in \mathbf{H}_N}{\text{argmin}}\ CV(\lambda)$$

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 (2009) 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.

5 Model Averaging and Endogeneity

The methods described above are all based on the strict exogeneity assumption of the regressors. This assumption implies that there is no correlation between the Xs and the unobservables (ϵ) affecting the output Y (i.e. $cov(X, \epsilon) = 0$). In the treatment effects terminology, it corresponds to the assumption that the treatment is conditionally randomly assigned to the population so that the ordinary least squares (OLS) estimates of the parameters can be interpreted as causal effects. However, in many applications such as empirical growth regressions this assumption is clearly violated. Therefore we might have some Xs, 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 $cov(X_1, \epsilon|X_2) \neq 0$). 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 $cov(Z, \epsilon|X_2) = 0$. 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.

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:

$$\begin{aligned} y &= X_1\beta_1 + X_2\beta_2 + \epsilon \\ X_1 &= Z\pi_1 + X_2\pi_2 + V \end{aligned} \tag{32}$$

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,[8] 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 labeled as reduced form equations).

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

$$U_i \sim N(0, \Sigma) \tag{33}$$

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 (pseudo) likelihood function for such a model and estimate the parameters via (pseudo) maximum likelihood (i.e. Limited Information Maximum Likelihood (LIML)), or we can estimate the parameters via two-stage least squares (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 above, 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

[8] 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.

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. 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 variables in X_1. This means that we must construct for this model the likelihood $f(y, X_1|X_2, Z)$ with the full set of candidate endogenous variables in order to guarantee comparability with all the other models under consideration, i.e., 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 (either X_1 for β_1, or X_2 for β_2 and π_2) not included in a particular model. 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.

In order to present the LIML likelihood, note that the model in (32) can be written as follows:

$$\left(\begin{array}{c|c} 1 & -\beta_1' \\ \hline 0 & I_{q_1} \end{array}\right) \left(\begin{array}{c} y' \\ \hline X_1' \end{array}\right) = \left(\begin{array}{c|c} \beta_2' & 0 \\ \hline \pi_2' & \pi_1' \end{array}\right) \left(\begin{array}{c} X_2' \\ \hline Z' \end{array}\right) + \left(\begin{array}{c} \epsilon' \\ \hline V' \end{array}\right) \tag{34}$$

or more compactly:

$$BY' = CW' + U' \tag{35}$$

where:

$$\begin{aligned} B &= \left(\begin{array}{c|c} 1 & -\beta_1' \\ \hline 0 & I_{q_1} \end{array}\right)_{Q\times Q} \\ Y' &= \left(\begin{array}{c} y' \\ \hline X_1' \end{array}\right)_{Q\times N} \\ C &= \left(\begin{array}{c|c} \beta_2' & 0 \\ \hline \pi_2' & \pi_1' \end{array}\right)_{Q\times(q_2+q_Z)} \\ W' &= \left(\begin{array}{c} X_2' \\ \hline Z' \end{array}\right)_{(q_2+q_Z)\times N} \\ U' &= \left(\begin{array}{c} \epsilon' \\ \hline V' \end{array}\right)_{Q\times N} \end{aligned}$$

The gaussian log-likelihood function of the full model (i.e. the model that includes all the candidate variables) is:

$$\ln f(Y|W, M_f) = -\frac{NQ}{2}\log 2\pi + N\log|\det B| - \frac{N}{2}\log\det\Sigma - \frac{1}{2}tr(\Sigma^{-1}U'U) \qquad (36)$$

where note that Y includes both y and X_1, W includes X_2 and Z, and M_f refers to the full model including all the candidate variables available. Despite the fact that number of parameters to be estimated might be huge and the problem might become computationally unfeasible from a model averaging perspective, we can concentrate out the reduced form parameters and drastically reduce the computational burden (see Moral-Benito (2010b)).

Having the likelihood function for the full model, it is easy to obtain the likelihood functions for the remaining models in order to compute the marginal likelihoods and model weights (or posterior model probabilities). Given the focus on model averaging, we need all the model-specific likelihoods in which some parameters are restricted to be zero depending on the variables (either endogenous or exogenous) included in the model. If a given endogenous variable is excluded from the full model, we simply restrict to zero the corresponding element of the β_1 vector of coefficients, but the rest of the likelihood remains unchanged in order to guarantee comparability across models. If the excluded variable is an exogenous one, we restrict to zero the corresponding elements of the vectors β_2 and π_2.[9]

As advanced by Raftery (1995), we can extend the model averaging approach to the setting of endogenous variables by computing BIC weights from the LIML likelihoods just described. Given the likelihood function, Section 3.4.2 of this survey formally presents the approach and its Bayesian justification. Moreover, this is also the approach considered by Moral-Benito (2010a) and Moral-Benito (2010b) in a panel data setting (see below for more details).

In the cross-sectional setting, Durlauf et al. (2008) represents the first attempt to address the issue of endogenous regressors in a BMA context.[10] More concretely, the paper is 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 $j = 1, ..., 2^{q_1+q_2}$. The authors propose to use 2SLS model-specific estimates for each

[9]Note that this particular likelihoods with restrictions would not be necessary if we are not interested in model comparison and our only interest is the estimation of a single model.

[10]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.

single model, and then take the average:

$$E(\theta|y) = \sum_{j=1}^{2^{q_1+q_2}} P\left(M_j|y\right) E(\theta|y, M_j) \approx \sum_{j=1}^{2^{q_1+q_2}} P\left(M_j|y\right) \widehat{\theta}^j_{2SLS} \tag{37}$$

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 $P\left(M_j|y\right) \propto f(y|M_j)P(M_j)$ are inspired in a limited information version of the BIC (i.e. LIBIC) approximation to the integrated likelihood $f(y|M_j)$:

$$f(y|M_j) \approx \exp\left[-\frac{N(q_1+1)}{2}\log(2\pi) - \frac{N}{2}\log(\det(\hat{\Sigma})) - \frac{h}{2}\log N\right] \tag{38}$$

where $\hat{\Sigma} = \sum_{i=1}^{N} \hat{U}_i\hat{U}_i'$ and $\hat{U}_i$ is the predicted residual from the 2SLS estimates. With respect to model priors ($P(M_j)$), Durlauf et al. (2008) use the dilution priors previously described.

The main drawback of this approach is that its formal justification remains an open issue as stated by the authors. They only give an heuristic interpretation to the results emerging from this averaging of 2SLS estimates. Moreover, since the model weights are based on "pseudo" likelihoods that are in principle not comparable across models, the comparability of these model weights represents another important caveat of the approach. This is so because in this approach, if an endogenous variable is not included in the model, its associated reduced form equations are not considered either. Then, the different models have "pseudo" likelihoods functions for different variables which are not comparable (e.g. the likelihood $f(y, x_1|z)$ is not comparable to the likelihood $f(y, x_2|z)$). More recently, Durlauf et al. (2009) consider model averaging across just-identified models so that model-specific 2SLS estimates $\widehat{\theta}^j_{2SLS}$ coincide with model-specific LIML estimates $\widehat{\theta}^j_{LIML}$ so that the proper likelihood-based BIC weights have formal justification but are still not comparable across models.

In a recent paper, Eicher et al. (2009b) 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 and thus we do not know which instruments to include in a given model. In the previous setting, the inclusion of a given endogenous variable in the model implied its own set of valid instruments to be included in the model. In particular, Eicher et al. (2009b) propose a 2-step procedure that first averages across the first-stage models (i.e. linear regressions of the endogenous variables on the exogenous

ones and 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. In this approach, some issues such as the selection of instruments (i.e. identification of some particular models which depends on q_1 and q_Z) and the statistical justification for the use of BIC-inspired weights in this context are still unclear.

All in all, given the potential criticisms of the approaches to model averaging and endogeneity in the existing literature (e.g. Durlauf et al. (2008), Eicher et al. (2009b)), the approach from first principles presented at the beginning of this section seems to be the most appropriate one to simultaneously address the issues of model uncertainty and endogeneity. The proposed approach, based on averaging across LIML estimates, guarantees comparability of the model likelihoods, and it has statistical justification (see Raftery (1995)). Its main disadvantage is the greater computational burden due to the large number of reduced form parameters to be estimated in all the models under consideration.[11]

5.1 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 ($i = 1, ..., N$) over different time periods ($t = 1, ..., T$). Therefore, the correlation between X and ϵ might arise because of a time-invariant and individual-specific characteristic (η_i), that is a component of ϵ_i ($\epsilon_i = \eta_i + \vartheta_{it}$) so that:

$$y_{it} = x_{it}\beta + \eta_i + \vartheta_{it} \tag{39}$$

where:

$$E(\eta_i|x_i) \neq 0 \tag{40}$$

$$E(\vartheta_{it}|x_i, \eta_i) = 0 \tag{41}$$

where x_i is a $T \times 1$ vector such that $x_i = (x_{i1}, x_{i2}, ..., x_{iT})'$.

Assumption (40) indicates that the regressors are correlated with the time invariant component of the error term and thus they are endogenous with respect to the fixed

[11] Note however that concentration of the likelihood functions with respect to the common parameters reduces this computational burden (e.g. Moral-Benito (2010b)).

effects. However, assumption (41) is usually labeled as a strict exogeneity assumption that represents the independence of the transitory shocks (ϑ) with respect to both the regressors and the permanent component of the shock (η).

Moral-Benito (2010a) considers such a panel setting and shows how to combine different panel data estimators with BMA techniques using different prior structures. In particular, Moral-Benito (2010a) extends to the panel data setting in (39)-(41) the Benchmark g-Priors and a diffuse prior (in the spirit of Sala-i-Martin et al. (2004)) on the parameter space, and the Binomial and Binomial-Beta priors on the model space. For the model weights, he considers BIC inspired weights that have both a Bayesian (g-UIP prior) and a Frequentist (Schwarz) interpretation.

In this framework we might also have dynamics that complicate the model-specific estimation step. In particular, the vector x_{it} can also include a lagged dependent variable (y_{it-1}) which is correlated with ϑ_{it-1} by definition and thus assumption (41) is violated. Moral-Benito (2010a) also considers a dynamic model as follows:

$$y_{it} = \alpha y_{it-1} + x_{it}\beta + \eta_i + \vartheta_{it} \tag{42}$$

where:

$$E(\eta_i|y_i, x_i) \neq 0 \tag{43}$$

$$E(\vartheta_{it}|y_i^{t-1}, x_i, \eta_i) = 0 \tag{44}$$

where y_i^{t-1} is a $(t-1) \times 1$ vector such that $y_i^{t-1} = (y_{i1}, y_{i2}, ..., y_{it-1})'$. Assumption (44) makes it clear that endogeneity generated by the dynamics of the model is taken into account.

In this setting, uncertainty comes from the selection of the xs to be included in the model. Moral-Benito (2010a) proposes to combine BMA with a panel likelihood-based estimator which allows obtaining consistent estimates of the autoregressive parameter α based on assumptions (43)-(44). Moreover, since model-specific estimates are based on a proper likelihood function, model weights are given by the BIC approximation to the marginal likelihood (see Raftery (1995)). The approach is labeled Bayesian Averaging of Maximum Likelihood Estimates (BAMLE).

On the other hand, panel data can also be useful for addressing the biases arising from reverse causality, which is a source of bias different from the biases just described. Coming back to the static setting in (39), the reverse causality problem arises if, instead

of the assumptions in (40)-(41), we face:

$$\begin{aligned} E(\eta_i|x_i) &\neq 0 \\ E(\vartheta_{it}|x_i,\eta_i) &\neq 0 \end{aligned} \tag{45}$$

In this setting, using panel data without additional instruments (Z) one can obtain causal effect estimates based on the following identification strategy: realizations of the endogenous regressors far enough in time are independent of the current shocks[12] (i.e. $cov(x_{it-\tau}, \vartheta_{it}) = 0$). Then, we can use this previous realizations ($x_{it-\tau}$) as "internal" instruments in the spirit of (32). Using this strategy, Moral-Benito (2010b) constructs a likelihood function for panel data models with unobserved heterogeneity (i.e. fixed effects) and endogenous regressors and combines this likelihood with BMA methods employing the BAMLE approach.

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). The formal justification of the considered model-specific "pseudo" likelihoods is an important caveat of this approach as in the cross-sectional case discussed in the previous subsection.[13]

6 Model Averaging in Economics

Until the 1990s, 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), Fernández et al. (2002), Volinsky et al. (1997)), and, on the other hand, papers from the forecasting combination literature to be discussed below. However, since the beginning of the 21st century, new methods together with more powerful computers are inspiring a flurry of BMA activity in different fields of economics.

Empirical growth is, without any doubt, the most active field in which model averaging techniques are being applied. In the search for a satisfactory empirical model of growth, the main area of effort has been the selection of appropriate variables to include in linear

[12]Note that this is the same strategy as the one adopted in the dynamic panel setting above in which one could interpret that the lagged dependent variable is an additional endogenous regressor.

[13]It is also worth mentioning the well-documented finite sample biases in the panel GMM estimates they consider (see for instance Alvarez and Arellano (2003)).

growth regressions. The literature concerned with this task is enormous: a huge number of papers have claimed to have found one or more variables correlated with the growth rate, resulting in a total of more than 140 variables proposed as growth determinants. However, given the limited number of observations, the fragility of these regressions causes a big concern among growth researchers.

In an attempt to investigate the robustness of the results, Levine and Renelt (1992) employ the extreme-bounds analysis proposed by Leamer (1983) and Leamer and Leonard (1983), and they concluded that very few variables (e.g. investment) were robustly correlated with growth. In contrast, Sala-i-Martin (1997) relaxed the robustness requirements, constructed weighted averages of OLS coefficients and found that some were fairly stable across specifications. However, the way in which standard errors and distribution of estimates are computed in Sala-i-Martin (1997) are somehow ad hoc and they lack formal justification. 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 in a pseudo-Bayesian approach denominated Bayesian Averaging of Classical Estimates (BACE). On the other hand, Fernández et al. (2001b) employ the Benchmark g Priors for the parameters in a pure Bayesian spirit. Both approaches use the Binomial prior on the model space with different prior expected model sizes (other papers applying these priors to the empirics of growth are Masanjala and Papageorgiou (2008) and Crespo-Cuaresma et al. (2009)). Magnus et al. (2010) consider the WALS approach with uniform model priors in the growth context, and Wagner and Hlouskova (2009) apply a FMA estimator based on principal components using four weighting schemes: equal, MMA, AIC, and BIC. Durlauf et al. (2008) is the first paper worried about causal effects in BMA empirical growth research using BIC weights and dilution priors on the model space. Moral-Benito (2010a) extends the use of model averaging techniques to a panel data setting considering and comparing different prior structures on both the parameter space and the model space. In the spirit of Raftery (1995), Moral-Benito (2010a) proposes to combine maximum likelihood estimates with model averaging using BIC weights in the so-called Bayesian Averaging of Maximum Likelihood Estimates (BAMLE) approach. Finally, Moral-Benito (2010b) and Mirestean and Tsangarides (2009) combine model averaging techniques with causal effect estimates in a panel data framework.

Since the seminal paper by Bates and Granger (1969), there has been an enormous literature on forecast combination with the aim of improving forecasting performance. From a Frequentist viewpoint, there is a vast empirical literature on forecast combining, and

there are also a number of simulation studies that compare the performance of combining methods in controlled experiments. These studies are surveyed by Diebold and Lopez (1996) and Timmermann (2006). With respect to the Bayesian approach to model averaging, there are many BMA applications to forecasting financial variables such as stock returns (e.g. Avramov (2002), Cremers (2002)) or exchange rates (e.g. Wright (2008a)). In the macro forecasting literature, Garratt et al. (2003) employ BMA to predict inflation and output growth in the UK and Wright (2008b) forecasts US inflation by BMA.

Aside from empirical growth and forecast combination, model averaging techniques are becoming popular in other fields of economics. Pesaran et al. (2009) employ model averaging as a remedy to the risk of inadvertently using false models in portfolio management; 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; Eicher et al. (2009a) study the effect of Preferential Trade Agreements (PTAs) on trade flows using BMA to account for the existent model uncertainty problem in this literature; 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, etc.) in 311 rural U.S. counties in the 1990's and 2000; in labor economics, Tobias and Li (2004) apply model averaging to estimate Mincer equations; Cohen-Cole et al. (2007) study the controversial issue of the deterrent effect of capital punishment employing a BMA approach; Galbraith and Hodgson (2009) analyze the determinants of the value of works of art using model averaging; in the health economics literature, Morales et al. (2006) characterize the dose-response relationship between an environmental exposure and adverse health outcomes using model averaging techniques.

7 Conclusions and Future Research

The choice of control variables in regression analysis applied to empirical research in economics is a critical issue that is, in general, underestimated. As illustrated by Leamer (1983) among others, conclusions from empirical studies may well depend on the controls included, so that 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. Both the Frequentist and the Bayesian approaches to model averaging have been summarized. Bayesian Model Averaging (BMA) involves the elicitation of model and parameter priors; Frequentist Model Averaging (FMA) requires to choose model weights and model-specific estimators. Several alternatives on both sides have been described in this paper. Moreover, an attempt to connect both approaches is made in Section 3.4.2.

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 in the conditional IV and panel settings has been summarized in this paper. Allowing for endogenous regressors in the FMA approach could be an interesting topic for future research.

In a recent paper, Angrist and Pischke (2010) argue that the rise of design-based approaches is the main responsible for the credibility revolution in empirical economics in the last three decades. The treatment effects literature has represented a huge progress in the estimation of more credible causal effects. For instance, randomized experiments and regression discontinuity can be extremely useful for that purpose. Matching estimators might also be very useful, but their identifying exogeneity assumption is conditional on a set of covariates (i.e. it is necessary to control for a group of regressors in order to guarantee the randomness of the treatment assignment). Provided that a set of conditioning (or control) variables is required for the validity of the approach, fragility of results to different conditioning sets can potentially be a cause of concern. Extending the model averaging apparatus to non-parametric matching (or other design-based approaches) might be a fruitful line for future research.

A Appendix

A.1 Asymptotic Theory of FMA Estimators

Suppose the density of the model in Section 4 is:

$$f_{true} = f(y, \beta, \gamma) = f(y, \beta_0, \gamma_0 + \delta/\sqrt{N})$$

where β is a parameter present in all models with β_0 its corresponding true value. γ is a vector around its true value γ_0 with perturbation $\delta/\sqrt{N}$. This setting is the local misspecification framework considered in Hjort and Claeskens (2003) for deriving the asymptotic results of FMA estimators. Let $\mu_{true} = \mu(f_{true})$ be the quantity of interest being $\mu(\cdot)$ a known function. The estimator of μ_{true} under the model M_j is given by:

$$\hat{\mu}_{M_j} = \mu(\hat{\beta}_{M_j}, \hat{\gamma}_{M_j}, \gamma_0, M_j^C)$$

where $\hat{\beta}_{M_j}$ and $\hat{\gamma}_{M_j}$ are maximum likelihood estimates and M_j^C is the complement of M_j.

Hjort and Claeskens (2003) analyzed the asymptotic properties of the FMA estimator:

$$\hat{\mu}_{FMA} = \sum_{j=1}^{2^q} \omega_{M_j} \hat{\mu}_{M_j}$$

where ω_{M_j} is a weight function for model M_j given $D_N = \hat{\delta}_{full}$, an estimator of δ under the model with all the q controls (i.e. the full model).

Let us introduce some notation. The score function is given by:

$$\begin{pmatrix} U(y) \\ V(y) \end{pmatrix} = \begin{pmatrix} \partial \log f(y, \beta_0, \gamma_0)/\partial\beta \\ \partial \log f(y, \beta_0, \gamma_0)/\partial\gamma \end{pmatrix}$$

with $(1+q) \times (1+q)$ variance matrix at $(\beta_0, \gamma_0')'$ given by:

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \qquad \text{and inverse} \qquad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}$$

The following theorem corresponds to Theorem 4.1 in Hjort and Claeskens (2003), and it provides the asymptotic distribution of the FMA estimator:

Theorem A.1 *If the weight functions ω_{M_j} sum to 1 and have at most a countable number of discontinuities, then:*

$$\sqrt{N}\,(\hat{\mu}_{FMA} - \mu_{true}) \xrightarrow{d} \Lambda = \mu_\beta' J_{00}^{-1} \zeta + w'[\delta - \hat{\delta}(D)]$$

where $D \sim N(\delta, \Psi)$ *is the limit of* D_N, $\Psi = J^{11}$, $\mu_\beta = \frac{\partial \mu}{\partial \beta}$ *evaluated at the point* (β_0, γ_0'), $\zeta \sim N(0, J_{00})$ *independent of* D, $\hat{\delta}(D) = \left\{ \sum \omega_{M_j} \pi_{M_j}' (\pi_{M_j} \Psi^{-1} \pi_{M_j}')^{-1} \pi_{M_j} \right\} \Psi^{-1} D$, *and* $w = J_{10} J_{00}^{-1} \mu_\beta - \mu_\gamma$. *Finally, let* π_{M_j} *be the projection matrix mapping* δ *to* δ_j.

A.2 Markov Chain Monte Carlo Model Composition

The Markov Chain Monte Carlo Model Composition (MC^3) algorithm proposed by Madigan and York (1995) generates a stochastic process that moves through model space. The idea is to construct a Markov chain of models $M(i), i = 1, 2, ...$ with state space Ψ. If we simulate this Markov chain for $i = 1, ..., N$, then under certain regularity conditions, for any function $h(M_j)$ defined on Ψ, the average:

$$\widehat{H} = \frac{1}{N} \sum_{i=1}^{N} h\left(M\left(i\right)\right)$$

converges with probability 1 to $E\left(h\left(M\right)\right)$ as $N \to \infty$. For example, to compute (7) in this fashion, we set $h(M_j) = E(\beta | y, M_j)$.

To construct the Markov chain, we define a neighborhood $nbd(M)$ for each $M \in \Psi$ that consists of the model M itself and the set of models with either one variable more or one variable fewer than M. Then, a transition matrix $\mathbf{q}$ is defined by setting $\mathbf{q}(M \to M') = 0$ $\forall\ M' \notin nbd(M)$ and $\mathbf{q}(M \to M')$ constant for all $M' \in nbd(M)$. If the chain is currently in state M, then we proceed by drawing M' from $\mathbf{q}(M \to M')$. It is the accepted with probability:

$$\min \left\{ 1, \frac{\Pr\left(M' | y\right)}{\Pr\left(M | y\right)} \right\}$$

Otherwise, the chain stays in state M.[14]

[14] Koop (2003) is a good reference for the reader interested in developing a deeper understanding of the MC^3 algorithm.

References

Alvarez, J. and M. Arellano (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica*, 71, 1121–1159.

Angrist, J. and J. Pischke (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics," *The Journal of Economic Perspectives*, 24, 3–30.

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 (General)*, 126, 255–258.

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

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

Buckland, S., K. Burnham, and N. Augustin (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.

Chen, H., A. Mirestean, and C. Tsangarides (2009): "Limited Information Bayesian Model Averaging for Dynamic Panels with Short Time Periods," *IMF Working Paper WP/09/74*.

Claeskens, G. and N. Hjort (2003): "The Focused Information Criterion," *Journal of the American Statistical Association*, 98, 900–916.

——— (2008): *Model Selection and Model Averaging*, Cambirdge University Press.

Clemen, R. (1989): "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559–583.

Cohen-Cole, E., S. Durlauf, J. Fagan, and D. Nagin (2007): "Model Uncertainty and the Deterrent Effect of Capital Punishment," *Federal Reserve Bank of Boston Working Paper*.

Cremers, K. (2002): "Stock Return Predictability: A Bayesian Model Selection Perspective," *The Review of Financial Studies*, 15, 1223–1249.

Crespo-Cuaresma, J., G. Doppelhofer, and M. Feldkircher (2009): "The Determinants of Economic Growth in European Regions," *CESifo Working Paper Series*.

Crespo-Cuaresma, J. and T. Slacik (2009): "On the Determinants of Currency Crises: The Role of Model Uncertainty," *Journal of Macroeconomics*, 31, 621–632.

Diebold, F. X. and J. A. Lopez (1996): "Forecast Evaluation and Combination," *Handbook of Statistics*, 241–268.

Doppelhofer, G. and M. Weeks (2009a): "Jointness of Growth Determinants," *Journal of Applied Econometrics*, 24, 209–244.

——— (2009b): "Jointness of Growth Determinants: Reply to Comments by Rodney Strachan, Eduardo Ley and Mark F.J. Steel," *Journal of Applied Econometrics*, 24, 252–256.

Durlauf, S., A. Kourtellos, and C. Tan (2008): "Are Any Growth Theories Robust?" *Economic Journal*, 118, 329–346.

——— (2009): "Is God in the Details? A Reexamination of the Role of Religion in Economic Growth," *Discussion Papers Series, Tufts University*, No. 0613.

Edgerton, H. and L. Kolbe (1936): "The Method of Minimum Variation for the Combination of Criteria," *Psychometrika*, 1, 183–188.

Eicher, T., C. Henn, and C. Papageorgiou (2009a): "Trade Creation and Diversion Revisited: Accounting for Model Uncertainty and Natural Trading Partner Effects," *Journal of Applied Econometrics*, forthcoming.

Eicher, T., A. Lenkoski, and A. Raftery (2009b): "Bayesian Model Averaging and Endogeneity Under Model Uncertainty: An Application to Development Determinants," University of Washington Working Paper UWEC-2009-19.

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

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

——— (2001b): "Model Uncertainty in Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16, 563–576.

——— (2002): "Bayesian Modeling of Catch in a Northwest Atlantic Fishery," *Journal of the Royal Statistical Society, Series C*, 51, 257–280.

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

Furnival, G. and R. Wilson (1974): "Regression by Leaps and Bounds," *Technometrics*, 16, 499–511.

Galbraith, J. and D. Hodgson (2009): "Dimension Reduction and Model Averaging for Estimation of Artists' Age-Valuation Profiles," *CIRANO Working Paper*.

Garratt, A., K. Lee, H. Pesaran, and Y. Shin (2003): "Forecast Uncertainties in Macroeconomic Modeling: An Application to the U.K. Economy," *Journal of the American Statistical Association*, 98, 829–838.

Geisser, S. (1965): "A Bayes Approach for Combining Correlated Estimates," *Journal of the American Statistical Association*, 60, 602–607.

George, E. (1999): "Discussion of "Model Averaging and Model Search Strategies" by M. Clyde," in *Bayesian Statistics*, ed. by J. Bernardo, A. Berger, P. Dawid, and S. A., Oxford University Press.

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. and J. Racine (2009): "Jackknife Model Averaging," *Unpublished Manuscript*.

Hjort, N. and G. Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899.

Hoeting, J., D. Madigan, A. Raftery, and T. Volinsky (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.

Kass, R. and L. Wasserman (1995): "A Reference Bayesian Test for Nested Hypothesis with Large Samples," *Journal of the American Statistical Association*, 90, 928–934.

Kleibergen, F. and E. Zivot (2003): "Bayesian and Classical Approaches to Instrumental Variable Regression," *Journal of Econometrics*, 114, 29–72.

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

Kuersteiner, G. and R. Okui (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*, Courcier, Paris.

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

——— (1983): "Let's Take the Con Out of Econometrics," *American Economic Review*, 73, 31–43.

Leamer, E. and H. Leonard (1983): "Reporting the Fragility of Regression Estimates," *Review of Economics and Statistics*, 65, 306–317.

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

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

——— (2009a): "Comments on Jointness of Growth Determinants," *Journal of Applied Econometrics*, 24, 248–251.

——— (2009b): "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression," *Journal of Applied Econometrics*, 24, 651–674.

Madigan, D. and A. Raftery (1994): "Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546.

Madigan, D. and J. York (1995): "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

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

Masanjala, W. and C. Papageorgiou (2008): "Rough and Lonely Road to Prosperity: A Reexamination of the Sources of Growth in Africa using Bayesian Model Averaging," *Journal of Applied Econometrics*, 23, 671–682.

McAleer, M., A. Pagan, and P. Volker (1985): "What Will Take the Con Out of Econometrics?" *American Economic Review*, 75, 293–307.

Mirestean, A. and C. Tsangarides (2009): "Growth Determinants Revisited," *IMF Working Paper*, WP/09/268.

Moral-Benito, E. (2010a): "Determinants of Economic Growth: A Bayesian Panel Data Approach," *The Review of Economics and Statistics*, forthcoming.

——— (2010b): "Panel Growth Regressions with General Predetermined Variables: Likelihood-Based Estimation and Bayesian Averaging," *CEMFI WP No. 1006*.

Morales, K., J. Ibrahim, C. Chen, and L. Ryan (2006): "Bayesian Model Averaging With Applications to Benchmark Dose Estimation for Arsenic in Drinking Water," *Journal of the American Statistical Association*, 101, 9–17.

Pesaran, H., C. Schleicher, and P. Zaffaroni (2009): "Model Averaging in Risk Management with an Application to Futures Markets," *Journal of Empirical Finance*, 16, 280–305.

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

Sala-i-Martin, X. (1997): "I Just Ran Two Million Regressions," *The American Economic Review*, 87, 178–183.

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

Stigler, S. (1973): "Laplace, Fisher, and the Discovery of the Concept of Sufficiency," *Biometrika*, 60, 439–445.

Strachan, R. (2009): "Comments on Jointness of Growth Determinants," *Journal of Applied Econometrics*, 24, 245–247.

Timmermann, A. (2006): "Forecast Combinations," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. Granger, and A. Timmermann, Amsterdam: North-Holland, 135–196.

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

Volinsky, C., D. Madigan, A. Raftery, and R. Kronmal (1997): "Bayesian Model Averaging in Proportional Hazard Models: Predicting the Risk of a Stroke," *Applied Statistics*, 46, 443–448.

Wagner, M. and J. Hlouskova (2009): "Growth Regressions, Principal Components and Frequentist Model Averaging," *Working Paper, Institute for Advanced Studies, Vienna.*

Wan, A. and X. Zhang (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.

——— (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 *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, ed. by P. Goel and A. Zellner, Amsterdam: North-Holland/Elsevier.