



WP/04/68

IMF Working Paper

A Bayesian Approach to Model Uncertainty

Charalambos G. Tsangarides

IMF Working Paper

African Department

A Bayesian Approach to Model Uncertainty

Prepared by Charalambos G. Tsangarides¹

Authorized for distribution by Lelde Schmitz

April 2004

Abstract

This Working Paper should not be reported as representing the views of the IMF.

The views expressed in this Working Paper are those of the author(s) and do not necessarily represent those of the IMF or IMF policy. Policy Discussion Papers describe research in progress by the author(s) and are published to elicit comments and to further debate.

This paper develops the theoretical background for the Limited Information Bayesian Model Averaging (LIBMA). The proposed approach accounts for model uncertainty by averaging over all possible combinations of predictors when making inferences about the variables of interest, and it simultaneously addresses the biases associated with endogenous and omitted variables by incorporating a panel data systems Generalized Method of Moments estimator. Practical applications of the developed methodology are discussed, including testing for the robustness of explanatory variables in the analyses of the determinants of economic growth and poverty.

JEL Classification Numbers: C11, C33

Keywords: Dynamic Panel Estimation, Bayesian Model Averaging

Author's E-Mail Address: ctsangarides@imf.org

¹This paper is a revised version of the first essay of my Ph.D. dissertation. I thank Mike Bradley, Bob Phillips, and Fred Joutz for their guidance, advice and support. Any remaining errors are my responsibility. Financial support from the Economic Club of Washington is gratefully acknowledged.

	Contents	Page
I.	Introduction	3
II.	Preliminaries on GMM	4
III.	GMM in the Bayesian Framework	7
	A. The Bayes' Estimator and GMM	7
	B. Limited Information Likelihood and GMM	9
IV.	Model Uncertainty and BMA	11
	A. Bayesian Hypothesis Testing	11
	B. The Information Criteria: BIC and LIBIC	12
	C. BMA under Limited Information	15
V.	Statistics for the Robustness Analysis	17
VI.	Conclusion	19
	References	20

I. INTRODUCTION

The study of socioeconomic phenomena may be plagued by inconsistent empirical estimates and model uncertainty. The case of inconsistent empirical estimates typically arises with omitted country-specific effects that, if not uncorrelated with other regressors, lead to a misspecification of the underlying dynamic structure, or with endogenous variables that may be incorrectly treated as exogenous. A panel data estimator that simultaneously addresses the issues of endogeneity and omitted variable bias is the systems Generalized Method of Moments Estimator (GMM) proposed by Hansen (1982). GMM estimators hold the potential for both consistency and efficiency gains by exploiting additional moment restrictions. The systems GMM involves the estimation of two equations, one in levels and the other in differences. The estimates from the difference equation, constructed by taking first differences of the levels equation eliminates the country-specific effect. For both equations, potentially endogenous explanatory variables are instrumented with their own lagged values, a factor that deals with the issue of endogeneity. Estimating the equations as a system, the procedure constrains similar coefficients to be constant across equations.²

The case of model uncertainty arises because the lack of clear theoretical guidance on the choice of regressors results in a wide set of possible specifications and, often, contradictory conclusions. Remedially, the analyst has three options: (i) arbitrarily select one model as the true model generating the data; (ii) present the results based on all plausible models without selecting between different specifications; and (iii) explicitly account for model uncertainty. While preferable, option (iii) presents enormous challenges at the level of both concept and statistical theory. Option (ii), although unsystematic, is preferable over option (i), but poses substantial logistical challenges. In practice, researchers tend to focus on one “channel” and choose option (i), ignoring model uncertainty altogether and risking overconfident inferences.³ In theory, accounting for model uncertainty requires some version of a “robustness check,” essentially an attempt to account for all possible combinations of predictors. A conceptually attractive solution to the problem of model uncertainty is provided by Bayesian Model Averaging (BMA) although difficulties at the implementation stage sometimes render it impractical.⁴ In particular, with a large number of regressors, k^* , the procedure may be infeasible due to the large number of models to be estimated, 2^{k^*} . In addition, the researcher is required to specify the prior distributions of all relevant parameters. In practice, most applications of BMA utilize an arbitrary set of priors, without examining the impact of this choice. Standard Bayesian Model Averaging techniques have been used in the context of investigating growth determinants

²To the extent that the lagged values of the regressors are valid instruments, this GMM estimator addresses consistently and efficiently both sources of bias.

³See Leamer (1978) and Raftery (1988), and (1996) for a discussion.

⁴Madigan and Raftery (1994) show that BMA provides optimal predictive ability. Hoeting and others (1999) summarize recent work using BMA. Brock and Durlauf (2000) provide an accessible explanation of criticisms levied at growth empirics and the contribution of Bayesian analysis in dealing with model uncertainty.

by Brock and Durlauf (2001), Doppelhofer, Miller, and Sala-i-Martin (2000) in their Bayesian Averaging of Classical Estimates (BACE) approach, and Fernandez, Ley, and Steel (2001).

Taking into consideration the concerns over model uncertainty, this essay develops the theory of a new Limited Information Bayesian Model Averaging estimator (LIBMA). The proposed estimator incorporates a dynamic panel estimator in the context of GMM, and a Bayesian robustness check to explicitly account for model uncertainty in evaluating the results of a universe of models generated by a set of possible regressors. The LIBMA approach provides certain advantages over the existing literature by relaxing the otherwise restrictive underlying assumptions in two ways. First, while standard Bayesian Model Averaging is a full information technique where a complete stochastic specification is assumed, LIBMA is a limited information approach that relies on GMM, a limited information technique based on moment restrictions rather than a complete stochastic specification. Second, while previous literature implicitly assumes exogenous regressors, LIBMA can control for endogeneity through the use of GMM.

The remainder of the paper is organized as follows. Section II introduces some preliminary ideas about the GMM. Section III constructs the GMM estimator in the Bayesian framework and the limited information likelihood. Section IV discusses the concepts of hypothesis testing and model selection in the Bayesian framework, presents the Limited Information Bayesian Information Criterion used in the context of GMM, and completes the derivation of the LIBMA. Section V presents all the calculated quantities and summary statistics on which the robustness analysis is based. The final section concludes.

II. PRELIMINARIES ON GMM

The GMM was developed by Hansen (1982) and White (1982) as an extension to the classical method of moments estimator. The basic idea of the GMM is to choose parameters of the model so as to match the moments of the model to those of the data as closely as possible. A weighing matrix determines the relative importance of matching each moment. Most common estimation procedures are contained in the GMM framework, including ordinary least squares, instrumental variables estimators, and in some cases, maximum likelihood estimators.

A key advantage to GMM over other estimation procedures is that there is no need to specify a likelihood function. The method of moments (and by extension, GMM) does not require the complete specification of distributions. Given that economic models do not specify joint distributions of economic variables, the method of moments (as well as other limited information inference methods) becomes very appealing in empirical studies. Of course, nothing comes for free. The cost is a loss of efficiency over methods such as Maximum Likelihood (MLE). The MLE can be viewed as a limiting case of GMM where under MLE the distribution of errors is

specified (so in a sense all of the moments are incorporated). The trouble with MLE is often that the errors may not follow a known distribution (such as the normal which is almost the universal standard in MLE).⁵ Thus, GMM offers a compromise between the efficiency of MLE and robustness to deviations from normality (or other distributional forms).

This section follows the presentation in Kim (2000) and (2002) to introduce the GMM concepts. Let x_t be an $n \times 1$ vector of stochastic processes defined on a probability space (Ω, \mathcal{F}, P) . Denote by $\mathbf{x}_T(\bar{\omega}) = (x_1(\bar{\omega}), \dots, x_T(\bar{\omega}))$, for $\bar{\omega} \in \Omega$, a T -segment of a particular realization of $\{x_t\}$. Let θ be a $q \times 1$ vector of parameters from $\Theta \subset \mathbb{R}^q$. Let \mathcal{G} be the Borel σ -algebra of Θ , where (Θ, \mathcal{G}) is a measurable space.⁶ In this paper Θ is a “grand” parameter space on which all the likelihoods, priors and posteriors under consideration are defined.

Let $h(x_t, \theta)$ be an $r \times 1$ vector valued function, $h : (\mathbb{R}^n \times \mathbb{R}^q) \rightarrow \mathbb{R}^r$. The function $h(x_t, \theta)$ characterizes an econometric relation $h(x_t, \theta_0) = w_t$ for a $\theta_0 \in \Theta$, where w_t is an r -vector stochastic disturbance process satisfying the standard conditions in GMM of Hansen (1982).

Assumption (A1)

$\{w_t, -\infty < t < \infty\}$ is stationary and ergodic.

Assumption (A2)

- (a) $E_P[w_t w_t']$ exists and is finite, and
- (b) $E_P[w_{t+s} | w_t, w_{t-1}, \dots]$ converges in mean square to zero.

Assumptions (A1) and (A2) imply a broad class of models as shown in Hansen (1982). Using iterated expectations, Assumption (A2) implies the $r \times 1$ moment conditions

$$E_P[h(x_t, \theta_0)] = 0 \tag{1}$$

Assumption (A3)⁷

- (a) $h(x, \cdot)$ is continuously differentiable in Θ for each $x \in \mathbb{R}^n$.
- (b) $h(\cdot, \theta)$ and $\partial h(\cdot, \theta) / \partial \theta$ are Borel measurable for each $\theta \in \Theta$.

Let $g_T(\mathbf{x}_T, \theta)$ be the sample average of $h(x_t, \theta)$ where $g_T(\mathbf{x}_T, \theta) \equiv \frac{1}{T} \sum_{t=1}^T h(x_t, \theta)$

⁵In such a case, one may use quasi-Maximum Likelihood estimation which does not sacrifice consistency. However, consistency may be an issue for nonlinear models estimated with Maximum Likelihood.

⁶The measurable space (Θ, \mathcal{G}) is required to define the posterior density in the Bayesian framework, and it is discussed here for completeness.

⁷Assumption (A3) is described here for completeness although it is used later for Lemma 1 as well as the derivation of the asymptotic normality of the posterior.

Definition 1 The GMM estimator $\left\{ \hat{\theta}_{G,T}(\omega) : T \geq 1 \right\}$ for some $\omega \in \Omega$ is the value of θ that minimizes the objective function

$$g_T(\mathbf{x}_T, \theta)' W_T^G g_T(\mathbf{x}_T, \theta), \quad (2)$$

where $\{W_T^G\}_{t=1}^\infty$ is a sequence of $(r \times r)$ positive definite weighting matrices which are functions of the data \mathbf{x}_T .

Assuming an interior optimum, the GMM estimate $\hat{\theta}_G$ is then the solution to the system of nonlinear equations:

$$\left\{ \frac{\partial g_T(\mathbf{x}_T, \theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right\}' \times W_T^G \times [g_T(\mathbf{x}_T, \hat{\theta})] = 0. \quad (3)$$

Let $R_w(s) = E_p[w_{s+1}w_1']$. Using Assumptions (A1) and (A2), it is ensured that $S = \sum_{s=-\infty}^\infty R_w(s)$ is well defined and finite. The matrix S above is sometimes interpreted as a long-run variance of $w_t = h(x_t, \theta_0)$ and can be alternatively written as

$$S \equiv \sum_{\nu=-\infty}^\infty E_P[h(x_t, \theta_0)h(x_{t-\nu}, \theta_0)']. \quad (4)$$

Conditions (1) and (4) form conditions on the first and second moments of $w_t = h(x_t, \theta_0)$ implied by the probability measure P . The matrix S is the asymptotic variance of $\sqrt{T}g_T(\mathbf{x}_T, \theta_0)$

$$S = \lim_{T \nearrow \infty} E_P[Tg_T(\mathbf{x}, \theta_0)g_T(\mathbf{x}, \theta_0)']. \quad (5)$$

In order to see how the weighing matrix in (2) works, consider first the situation where there are as many moment conditions as parameters (referred to as the “just-identified” case). The moments will all be perfectly matched and the objective function in (2) will have a value of zero. In the “over-identified” case where there are more moment conditions than parameters, not all of the moment restrictions will be satisfied, so the weighting matrix W_T^G determines the relative importance of the various moment conditions.

Hansen (1982) points out that setting $W_T^G = S^{-1}$, the inverse of an asymptotic covariance matrix, is optimal in the sense that it yields parameter estimates with the smallest asymptotic variance. (Intuitively, more weight is given to the moment conditions with less uncertainty. S is also known as the spectral density matrix evaluated at frequency zero.) There are many approaches for estimating (consistent estimators of) S which can account for various forms of heteroskedasticity and/or serial correlation, including White (1980), the Bartlett kernel used by Newey and

West (1987), the truncated kernel of Hansen (1982), and the automatic bandwidth selection from Andrews and Monahan (1992).

Let \widehat{S}_T be a consistent estimator of S based on a sample of size T . An optimal GMM estimator is obtained with $W_T^G = \widehat{S}_T^{-1}$ in (2)

$$g_T(\mathbf{x}_T, \theta)' \widehat{S}_T^{-1} g_T(\mathbf{x}_T, \theta), \quad (6)$$

where S is approximated by \widehat{S}_T

$$\widehat{S}_T = T g_T(\mathbf{x}, \theta_0) g_T(\mathbf{x}, \theta_0)'. \quad (7)$$

III. GMM IN THE BAYESIAN FRAMEWORK

In contrast to the classical approach, Bayesian estimation requires the specification of likelihood functions or the data generating mechanism. Because of this reason, one may conclude that the Bayesian method cannot be applied to the moment problem. However, recent developments in the Bayesian and classical econometrics have made it possible to consider a likelihood interpretation of some non-likelihood problems.⁸ Innovative work in this area was done by Zellner (1996 and 1997) who developed a finite sample Bayesian Method of Moments (BMOM) based on the principle of maximum entropy.⁹ One of the distinguishing features of the BMOM approach is that it yields post-data densities for models' parameters without use of an assumed likelihood function. Inoue (2001) proposes a semi-parametric Bayesian method of moments approach (which differs from the maximum entropy approach) that enables direct Bayesian inference in the method of moments framework. It turns out that the posterior distribution of strongly identified parameters is asymptotically normal even in the presence of weakly identified parameters. Finally, Kim (2000) and (2002) develops a limited information procedure in the Bayesian framework that does not require the knowledge of the likelihood function. His procedure is the Bayesian counterpart of the classical GMM but has certain advantages over the classical GMM for practical applications, and it is the approach we closely follow in this essay.

A. The Bayes Estimator and GMM

We now begin the construction of the GMM in the Bayesian framework. In the classical framework, GMM is a limited information procedure. The GMM estimate

⁸Inoue (2001), and Kim (2000) and (2001) provide a good literature review.

⁹As Golan, Judge, and Miller (1996) show, in the entropy approach, estimators are chosen to maximize entropy or minimize some distance metric between the true probability measure and artificial probability measures for which the moment condition in question is satisfied. Hence, this approach does not require knowledge of the functional form of the likelihood. Using the Bayesian counterpart of this approach, one can obtain finite-sample post-data moments and distribution of the parameters and conduct post-data inference (e.g. see Zellner 1996 and 1997). Many traditional estimators are special cases of entropy.

in Definition (1) is based on the moment condition (1), a set of limited information on the data generating process. The goal is to build a Bayesian counterpart of the classical GMM by constructing a Bayesian limited information procedure based on a set of moments.

Following Kim (2000) and (2002), we begin with some of the basic elements. A Bayesian framework is identified by a posterior density defined in the measurable space (Θ, \mathcal{G}) . Let $\pi_T(\theta|\mathbf{x}_T(\omega))$ be the “true” posterior of θ that may be unknown.¹⁰ Assume that the posterior $\pi_T(\cdot|\mathbf{x}_T(\cdot))$ is jointly measurable $\mathcal{G} \times \mathcal{F}$. Define

$$P_T^\pi(G, \omega) = \int_G \pi_T(\theta|\mathbf{x}_T(\omega))d\theta$$

for any $G \in \mathcal{G}$ and $\omega \in \Omega$, where $P_T^\pi(\cdot, \omega)$ is a probability measure on Θ for every $\omega \in \Omega$.

Let $\ell(\theta, \delta)$ be the loss function that penalizes for the choice of δ when θ is the real parameter value. The Bayes’ estimator is an estimator that minimizes the expected posterior loss

$$\hat{\theta}_B = \delta^*(\mathbf{x}_T) = \arg \min_{\delta} E^\pi[\ell(\theta, \delta)], \quad (8)$$

where

$$E^\pi[\ell(\theta, \delta)] = \int_{\Theta} \ell(\theta, \delta)\pi_T(\theta|\mathbf{x}_T)d\theta. \quad (9)$$

We are interested in a loss function that yields an estimator equivalent to the GMM estimator. Since the objective is to study a Bayesian counterpart of the classical GMM, it is natural to adopt a loss function with this property. Consider the following loss function that is quadratic in g_T :

$$\ell(\theta, \delta) = \mathcal{L}(g_T(\theta), g_T(\delta)) = [g_T(\theta) - g_T(\delta)]'W_T[g_T(\theta) - g_T(\delta)], \quad (10)$$

where $\{W_T^G\}_{t=1}^\infty$ is a sequence of positive definite weighing matrices. The loss function in (10) can be transformed to a loss function quadratic in θ :

$$\ell(\theta, \delta) = [\theta - \delta]'\widetilde{W}_T[\theta - \delta], \quad (11)$$

where $\widetilde{W}_T = \{\frac{\partial g(\tilde{\theta})}{\partial \theta}\}'W_T\{\frac{\partial g(\tilde{\theta})}{\partial \theta}\}$, and $\tilde{\theta} \in (\theta, \delta)$.

The loss functions in (10) and (11) are such that under some conditions (discussed in Lemma 1 below) yield an estimator that is the same as the GMM. As discussed in Kim (2000), the choice of the loss functions (10) and (11) does not cause loss of generality. The main results of this essay do not change so far as the chosen loss function can be transformed into a function that is quadratic in θ .

¹⁰It is the “true” posterior in the sense that it is obtained from the true likelihood of θ , or it is a posterior of θ that contains a richer set of information than that in the limited information posterior discussed in this paper.

From the minimization problem (8) using the loss function in (10) the first order condition is

$$E^\pi \left[\left\{ \frac{\partial g_T(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right\}' W_T g_T(\theta) \right] = \left\{ \frac{\partial g_T(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right\}' W_T g_T(\hat{\theta}). \quad (12)$$

This implies a moment condition

$$E^\pi[\Xi_T g_T(\theta)] = \Xi_T g_T(\hat{\theta}),$$

where the right hand side is a constant conditional on \mathbf{x}_T and

$$\Xi_T = \left\{ \frac{\partial g_T(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right\}' \times W_T.$$

Interpreting the GMM estimator as a Bayes' estimator, the right hand side of (12) is equal to zero. So we have the moment condition:

$$E^\pi[\Xi_T g_T(\theta)] = 0.$$

LEMMA 1:

Assume the second order conditions hold for the minimization in the GMM estimate in (2) and in the Bayes' estimate in (8) with the loss function described in (10). Then, under Assumption (A3), the GMM estimator $\hat{\theta}_{GMM}$ is equal to the Bayes' estimator $\hat{\theta}_B$ if and only if and only if

$$E^\pi[\Xi_T g_T(\theta)] = 0,$$

where

$$\Xi_T = \left\{ \frac{\partial g_T(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right\}' \times W_T$$

and $\{W_T\} = \{W_T^G\}$.¹¹

B. Limited Information Likelihood and GMM

In this section we follow the discussion in Section 3 of Kim (2002) to establish a semi-parametric limited information likelihood based on the moment conditions which form a set of limited information on the data generating mechanism. (This limited information likelihood is then used to derive a limited information posterior in Kim (2002).) The approach to get this limited information likelihood function is based on the principle of maximum entropy where the idea is to get a likelihood that is closest to the unknown true likelihood in an information distance.

¹¹The proof of this lemma is found in Kim (2001).

From the moment condition (1) we have

$$E_P[g_T(\mathbf{x}_T, \theta_0)] = 0, \quad (13)$$

and from (4) and (6) we have the second moment condition on g_T

$$S = \lim_{T \nearrow \infty} E_P[Tg_T(\mathbf{x}_T, \theta_0)g_T(\mathbf{x}_T, \theta_0)'], \quad (14)$$

where S is the long-run variance of $w_t = h(x_t, \theta_0)$ described in (4). Under Assumptions (A1) and (A2) it can be shown¹² that

$$\lim_{T \nearrow \infty} E_P[Tg_T(\mathbf{x}_T, \theta_0)'S_T^{-1}g_T(\mathbf{x}_T, \theta_0)'] = r. \quad (15)$$

Given the true probability measure P with the properties in the moment conditions (13) and (14), we are interested in the probability measure Q that implies the same moment conditions. Let \mathcal{Q} be a family of probability measures that is absolutely continuous with respect to P such that for $\theta \in \Theta$

$$\mathcal{Q}(\theta) = \{Q : E_Q[g_T(\mathbf{x}_T, \theta)] = 0\} \cap \{Q : \lim_{T \nearrow \infty} E_Q[Tg_T(\mathbf{x}_T, \theta)g_T(\mathbf{x}_T, \theta)'] = S\} \quad (16)$$

which (as shown in Kim 2002) reduces to

$$\mathcal{Q}(\theta) = \{Q : \lim_{T \nearrow \infty} E_Q[Tg_T(\mathbf{x}_T, \theta)'S_T^{-1}g_T(\mathbf{x}_T, \theta)] = r\}. \quad (17)$$

Usually such a Q is not unique. For $Q \in \mathcal{Q}$ we are interested in the one that it is the closest to the true probability measure P in the entropy distance or the Kullback-Leibler information distance (White 1982) or the I -divergence distance (Csiszar 1975). The optimization problem yields such a solution Q^*

$$Q^*(\theta) = \arg \min_{Q \in \mathcal{Q}} I(Q||P) \equiv \int \ln(dQ/dP)dQ, \quad (18)$$

where dQ/dP is the Radon-Nikodym derivative (or density) of Q with respect to P . So, Q^* is the solution of the constrained minimization where the constraint is given with respect to the moments implied in the measure P . As in Csiszar (1975), Q^* is defined to be the I -projection of P on \mathcal{Q} . Further, denote by $q_P^*(\theta) = dQ^*(\theta)/dP$ the Radon-Nikodym derivative of $Q^*(\theta)$ with respect to P . Kim (2002) calls $q_P^*(\theta)$ a *limited information density* or the I -projection density (following Csiszar (1975)).

The solution of (18) $q_P^*(\theta)$ is uniform in $\theta \in \Theta$ that satisfies the moment in (16) or (17), and therefore, $q_P^*(\theta)$ can be interpreted as the likelihood of θ . Thus, we call $q_P^*(\theta)$ a *Limited Information Likelihood* (LIL) or the I -projection likelihood.

Under the conditions on Q

¹²The proof is discussed in Kim (2002).

$$q_P^*(\theta) = \mathcal{K} \exp\left\{\lim_{T \nearrow \infty} \kappa T g_T(\mathbf{x}_T, \theta)' S_T^{-1} g_T(\mathbf{x}_T, \theta)\right\}, \quad (19)$$

where κ is a constant and \mathcal{K} is a normalizing constant. As shown in Kim (2002), $\kappa = -1/2$ is a desirable choice. Finally, Theorem 1 of Kim (2002) establishes that $q_{P,T}^*(\mathbf{x}_T, \theta)$ is a finite sample analogue of $q_P^*(\theta)$. Therefore, the sample LIL is

$$q_{P,T}(\mathbf{x}_T, \theta) = \mathcal{K} \exp\left\{\kappa T g_T(\mathbf{x}_T, \theta)' S_T^{-1} g_T(\mathbf{x}_T, \theta)\right\}. \quad (20)$$

When S is not known it is replaced by a consistent estimator \widehat{S}_T .

IV. MODEL UNCERTAINTY AND BMA

Standard statistical practice ignores model uncertainty. The classical approach conditions on a single model and thus leads to underestimation of uncertainty when making inferences about quantities of interest. A complete Bayesian solution to the problem of model uncertainty is the BMA approach which involves averaging over all possible combinations of predictors when making inferences about the quantities of interest.¹³ The Bayesian approach avoids conditioning on a single model. No model is assumed to be the “true” model, instead, all possible models are assigned different probabilities based on the researcher’s prior beliefs using the posterior model probabilities as weights. As noted by Hoeting and others (1994), this is reasonable as it allows for propagation of model uncertainty into the posterior distribution and leads to more sensible uncertainty bands.

The following sections draw from Raftery (1994), Kass and Raftery (1995), and Kim (2000) and (2002). First, we introduce Bayesian hypothesis testing and Bayes factors to test competing models. Then, we derive a limited information model selection criterion, in order to calculate the Bayes factors in the case of a limited information procedure. Finally, we incorporate the derived criterion in the context of BMA to derive the posterior distributions of the parameters of interest.

A. Bayesian Hypothesis Testing

We begin with the general setup for a model selection problem. Let \mathcal{M} be a family of candidate models for \mathbf{x}_T . A model $M_k \in \mathcal{M}$ is associated with a parameter space Θ^k of dimension q_k for $k \in \mathbf{I}$ where $\mathbf{I} = \{1, \dots, I\}$ and characterized by a relation of the form $h(x_t, \theta) = w_t$ (as described in Section 1) with w_t a stochastic process satisfying Assumptions (A1) and (A2). For every $M_k \in \mathcal{M}$ a set of moment conditions is defined as in (1) and a likelihood $q_T^k(\mathbf{x}_T, \theta_k)$ is defined (as in (20)).

¹³Madigan and Raftery (1994) note that the BMA approach provides the optimal predictive ability. Hoeting and others (1999) summarize recent work using BMA.

Suppose that we want to use data \mathbf{x}_T to test competing hypotheses presented by two models M_1 and M_2 with parameter vectors θ_1 and θ_2 . Let $p(M_1|\mathbf{x}_T)$ be the posterior probability that M_1 is the correct model,

$$p(M_1|\mathbf{x}_T) = \frac{q_T(\mathbf{x}_T|M_1)p(M_1)}{q_T(\mathbf{x}_T|M_1)p(M_1) + q_T(\mathbf{x}_T|M_2)p(M_2)}, \quad (21)$$

where (for $k = 1, 2$) $q_T(\mathbf{x}_T|M_k)$ is the marginal probability of the data given M_k , and $p(M_k)$ is the prior probability of model M_k .¹⁴

In general, the term $q_T(\mathbf{x}_T|M_k)$ in (21) is obtained by integrating over the parameter space

$$q_T(\mathbf{x}_T|M_k) = \int_{\Theta^i} q_T(\mathbf{x}_T|\theta_k, M_k)\phi(\theta_k|M_k)d\theta_k, \quad (22)$$

where $q_T(\mathbf{x}_T|\theta_k, M_k) = q_T^k(\mathbf{x}_T, \theta_k)$, the likelihood of θ_k under model M_k (the marginal likelihood), and $\phi(\theta_k|M_k)$ is the prior density associated with model M_k .

The posterior odds ratio for M_2 against M_1 (i.e. the ratio of their posterior probabilities $\frac{p(M_2|\mathbf{x}_T)}{p(M_1|\mathbf{x}_T)}$) can be used to measure the extent to which the data support M_2 over M_1 . Using (21) the posterior odds ratio is

$$\frac{p(M_2|\mathbf{x}_T)}{p(M_1|\mathbf{x}_T)} = \frac{q_T(\mathbf{x}_T|M_2)}{q_T(\mathbf{x}_T|M_1)} \times \frac{p(M_2)}{p(M_1)}, \quad (23)$$

where the first term on the right-hand side of (23) is the Bayes factor for M_2 against M_1 , denoted by B_{21} , and the second term is the prior odds ratio. Sometimes the prior odds ratio is set to 1, representing the lack of prior preference for either model, in which case the posterior odds ratio is equal to the Bayes factor. When the posterior odds ratio is greater (less) than 1, the data favor M_2 over M_1 (M_1 over M_2).

Evaluating the Bayes factor in (23) for hypothesis testing requires calculating the marginal likelihood $q_T(\mathbf{x}_T|M_k)$. This can be a high-dimensional and intractable integral. Various analytic and numerical approximations have been proposed which are reviewed in Kass and Raftery (1995). The Bayesian Information Criterion (BIC) is a simple and accurate method to estimate Bayes factors when the likelihood function is known. This is discussed first in the next section. Then we extend the discussion to the case where only the limited information likelihood is available to derive the Limited Information Bayesian Information Criterion (LIBIC).

B. The Information Criteria: BIC and LIBIC

Following the approach of Raftery (1994), we focus on approximating the marginal likelihood for a single model, that is, the right hand side of (22). We will avoid

¹⁴Similar to (21), $p(M_2|\mathbf{x}_T) = \frac{p(\mathbf{x}_T|M_2)p(M_2)}{p(\mathbf{x}_T|M_1)p(M_1)+p(\mathbf{x}_T|M_2)p(M_2)}$, and $p(M_1|\mathbf{x}_T) + p(M_2|\mathbf{x}_T) = 1$.

indexing for a specific model, so a general form of (22) can be written as $q_T(\mathbf{x}_T|M) = \int q_T(\mathbf{x}_T|\theta, M)\phi(\theta|M)d\theta$.

Let $f(\theta) = \log[q_T(\mathbf{x}_T|\theta, M)\phi(\theta|M)]$, and consider a Taylor series expansion of $f(\theta)$ about $\tilde{\theta}$, the value of θ that maximizes $f(\theta)$ or the posterior mode. The expansion gives

$$f(\theta) = f(\tilde{\theta}) + (\theta - \tilde{\theta})' f'(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})' f''(\tilde{\theta})(\theta - \tilde{\theta}) + o(\|\theta - \tilde{\theta}\|)^2 \quad (24)$$

where $f'(\theta) = (\frac{\partial f(\theta)}{\partial \theta_1}, \dots, \frac{\partial f(\theta)}{\partial \theta_d})$ is the vector of first partial derivatives of $f(\theta)$, and $f''(\theta)$ is the Hessian matrix of second partial derivatives of $f(\theta)$ whose (i, j) element is $\frac{\partial^2 f(\theta)}{\partial \theta_i \partial \theta_j}$. Since $f(\theta)$ is maximized at $\tilde{\theta}$, $f'(\tilde{\theta}) = 0$, so (24) becomes

$$f(\theta) \approx f(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})' f''(\tilde{\theta})(\theta - \tilde{\theta}). \quad (25)$$

From the definition of $f(\theta)$ and (22) it follows that $q_T(\mathbf{x}_T|M) = \int \exp[f(\theta)]d\theta$, and using (25):

$$q_T(\mathbf{x}_T|M) \approx \exp[f(\tilde{\theta})] \times \int \exp[\frac{1}{2}(\theta - \tilde{\theta})' f''(\tilde{\theta})(\theta - \tilde{\theta})]d\theta. \quad (26)$$

Recognizing that the integrand in (26) as proportional to a multivariate normal density gives

$$q_T(\mathbf{x}_T|M) \approx \exp[f(\tilde{\theta})](2\pi)^{\frac{d}{2}}|A|^{-\frac{1}{2}}, \quad (27)$$

where d is the number of parameters in the model and $A = -f''(\tilde{\theta})$. This is the Laplace approximation method.¹⁵ Using

$$\log q_T(\mathbf{x}_T|M) = \log q_T(\mathbf{x}_T|\tilde{\theta}, M) + \log \phi(\tilde{\theta}|M) + (\frac{d}{2}) \log(2\pi) - \frac{1}{2} \log |A| + O(n^{-1}). \quad (28)$$

In large samples, $\tilde{\theta} = \hat{\theta}$, where $\hat{\theta}$ is the Maximum Likelihood Estimator, and $A \approx n\mathbf{i}$ where \mathbf{i} is the expected Fisher information matrix for one observation. This is a $(d \times d)$ matrix whose (i, j) element is $-E[\frac{\partial^2 \log p(y_1|\theta)}{\partial \theta_i \partial \theta_j} |_{\theta=\hat{\theta}}]$, with the expectation being taken over values of y_1 with θ held fixed. Thus, $|A| \approx n^d |\mathbf{i}|$. With these approximations and an added $O(n^{-\frac{1}{2}})$ error, (28) becomes

$$\log q_T(\mathbf{x}_T|M) = \log q_T(\mathbf{x}_T|\hat{\theta}, M) + \log \phi(\hat{\theta}|M) + (\frac{d}{2}) \log(2\pi) - (\frac{d}{2}) \log n - \frac{1}{2} \log |\mathbf{i}| + O(n^{-\frac{1}{2}}). \quad (29)$$

¹⁵Tierney and Kadane (1986) show that the error in (27) is $O(n^{-1})$ so that $nO(n^{-1}) \rightarrow \text{constant}$ as $n \rightarrow \infty$.

Removing the terms of order $O(1)$ or less, gives¹⁶

$$\log q_T(\mathbf{x}_T|M) = \log q_T(\mathbf{x}_T|\hat{\theta}, M) - \left(\frac{d}{2}\right) \log n + O(1). \quad (30)$$

Equation (30) is the approximation on which the BIC is based and was first derived by Schwarz (1978). As suggested by Raftery (1994), although the $O(1)$ term suggests that the error does not vanish with an infinite amount of data, the error will tend towards zero as a proportion of $\log q_T(\mathbf{x}_T|M)$, which ensures that the error will not affect the conclusion reached given enough data. For a particular choice of prior, the error term is of much smaller magnitude. Suppose that the prior is a multivariate normal with mean $\hat{\theta}$ and variance matrix \mathbf{i}^{-1} . Under that choice of a prior, we have

$$\log \phi(\hat{\theta}|M) = -\left(\frac{d}{2}\right) \log(2\pi) + \frac{1}{2} \log |\mathbf{i}|. \quad (31)$$

Substituting (31) in (29), we get an expression for $\log q_T(\mathbf{x}_T|M)$ where the error term vanishes as $n \rightarrow \infty$

$$\log q_T(\mathbf{x}_T|M) = \log q_T(\mathbf{x}_T|\hat{\theta}, M) - \left(\frac{d}{2}\right) \log n + O(n^{-\frac{1}{2}}). \quad (32)$$

Using the approximation in (32) we can derive the Bayes factor $B_{21} = \frac{p(\mathbf{x}_T|M_2)}{p(\mathbf{x}_T|M_1)}$ in (23), such that:

$$\log B_{21} = \log q_T(\mathbf{x}_T|\hat{\theta}_2, M_2) - \log q_T(\mathbf{x}_T|\hat{\theta}_1, M_1) - \frac{(d_2 - d_1)}{2} \log n + O(n^{-\frac{1}{2}}). \quad (33)$$

As discussed in Kass and Raftery (1995), the expression in (33) is the Schwarz criterion (S) and as $n \rightarrow \infty$, $\frac{S - \log B_{21}}{\log B_{21}} \rightarrow 0$. (Based on this result, (33) can be viewed as an approximation to the Bayes factor B_{21} .) Twice the Schwarz criterion is the BIC or

$$BIC = 2 \log B_{21} = 2[\log q_T(\mathbf{x}_T|\hat{\theta}_2, M_2) - \log q_T(\mathbf{x}_T|\hat{\theta}_1, M_1)] - (d_2 - d_1) \log n + O(n^{-\frac{1}{2}}). \quad (34)$$

Exact calculation of equation (34) requires the knowledge of the likelihood function for each of the models. If M_1 is nested within M_2 (34) reduces to $2 \log B_{21} \approx \chi_{21}^2 - df_{21} \log n$, where χ_{21}^2 is the standard likelihood ratio test for testing M_1 against M_2 and $df_{21} = d_2 - d_1$ is the number of degrees of freedom.

The full-information likelihood function is not available in the context of GMM. Therefore, in order to calculate the BIC in (34) we need to rely on the LIL developed by Kim (2002) and discussed in Section 3. Using the LIL in (30) we can replace

¹⁶Note that in (29) the terms $\log p(\hat{\theta}|M)$, $\left(\frac{d}{2}\right) \log(2\pi)$, $-\frac{1}{2} \log |\mathbf{i}|$, and $O(n^{-\frac{1}{2}})$ are all of order $O(1)$ or less.

the likelihood functions on the right hand side of (34) to get a Limited Information Bayesian Information Criterion (LIBIC). First, using (20) (with S replaced by a consistent estimator \widehat{S}_T) to substitute for the log-likelihood in (34) we have

$$\log q_{P,T}(\mathbf{x}_T|\widehat{\theta}, M) = \log \mathcal{K} + \kappa T g_T(\mathbf{x}_T, \theta)' \widehat{S}_T^{-1} g_T(\mathbf{x}_T, \theta), \quad (35)$$

and substituting in (34) eliminating the $O(n^{-\frac{1}{2}})$ term we have the expression for the LIBIC:

$$LIBIC = 2[\log q_{P,T}(\mathbf{x}_T|\widehat{\theta}_2, M_2) - \log q_{P,T}(\mathbf{x}_T|\widehat{\theta}_1, M_1)] - (d_2 - d_1) \log n, \quad (36)$$

which reduces to

$$LIBIC = 2[\kappa T g_T(\mathbf{x}_T, \theta_2)' \widehat{S}_T^{-1} g_T(\mathbf{x}_T, \theta_2) - \kappa T g_T(\mathbf{x}_T, \theta_1)' \widehat{S}_T^{-1} g_T(\mathbf{x}_T, \theta_1)] - (d_2 - d_1) \log n. \quad (37)$$

The LIBIC expression in (37) can be calculated from the estimated output and will be used to estimate the Bayes factors required for the hypothesis testing.

C. BMA under Limited Information

Suppose we can divide the parameter space into K regions (models), so we have the space of all possible models $\mathcal{M} = \{M_j : j = 1, \dots, K\}$. Let Δ be the quantity of interest (such as a parameter, in our case). Then Bayesian inference about Δ is constructed given the data D , based on the law of total probability:

$$p(\Delta|D) = \sum_{k=1}^K p(\Delta|D, M_k) p(M_k|D), \quad (38)$$

where $p(\Delta|D)$, the posterior distribution of the quantity of interest Δ is a mixture of the posterior distributions of that quantity under each of the models with mixing probabilities given by the posterior model probabilities and using the posterior model probabilities as weights. Thus, the *full* posterior distribution of Δ is a weighted average of the posterior distributions under each model (M_1, \dots, M_K), where the weights are the posterior model probabilities $p(M_k|D)$. This procedure is what is typically referred to as BMA, and it is in fact the standard Bayesian solution under model uncertainty, since it follows from direct application of Bayes' theorem.

Denoting the data by \mathbf{x}_T , (38) becomes $p(\Delta|\mathbf{x}_T) = \sum_{k=1}^K p(\Delta|\mathbf{x}_T, M_k) p(M_k|\mathbf{x}_T)$. Using Bayes' theorem, the posterior model probabilities are obtained using (21) extended for the case of K models, such that:

$$p(M_k|\mathbf{x}_T) = \frac{q_T(\mathbf{x}_T|M_k) p(M_k)}{\sum_{j=1}^K q_T(\mathbf{x}_T|M_j) p(M_j)}. \quad (39)$$

Although this fully Bayesian approach is a feasible and a very attractive solution to the problem of accounting for model uncertainty, there are certain difficulties in the implementation of the BMA, making it (in some cases) a rather unpopular and less practical proposition. First, when the number of regressors k^* is very large, the number of models is 2^{k^*} and, as a result, calculations may be rendered infeasible.¹⁷ Second, the BMA requires specification of the prior distributions of all relevant parameters so for k^* possible regressors, 2^{k^*} priors are needed. In most BMA cases, the choice of priors has essentially been arbitrary and the impact of this choice on the estimated parameters has not been examined. In the context of growth Brock and Durlauf (2001), Doppelhofer, Miller, and Sala-i-Martin (2000), and Fernandez, Ley, and Steel (2001) are all applications of BMA techniques to investigate robustness of growth determinants in light of model uncertainty.¹⁸

Each of the K models is compared in turn with a baseline model M_0 (which could be the null model with no independent variables), yielding Bayes factors $B_{10}, B_{20}, \dots, B_{K0}$. The value of BIC for model M_k denoted BIC_k , is the approximation to $2 \log B_{0k}$ given by (34), where B_{0k} is the Bayes factor for model M_k against M_0 .¹⁹

It is possible to write (21) and (39) in terms of the BIC. To see this, rewrite the Bayes factor $B_{12} = \frac{p(D|M_1)}{p(D|M_2)}$ as $B_{12} = \frac{p(D|M_1)}{p(D|M_2)} \times \frac{p(D|M_0)}{p(D|M_0)} = \frac{\frac{p(D|M_1)}{p(D|M_0)}}{\frac{p(D|M_2)}{p(D|M_0)}} = \frac{B_{10}}{B_{20}} = \frac{B_{02}}{B_{01}}$. Using (34), this implies that $2 \log B_{12} = 2(\log B_{02} - \log B_{01}) = BIC_2 - BIC_1$. Substituting the Bayes factor B_{12} in (21), we get $p(M_1|\mathbf{x}_T) = \frac{B_{12}q_T(\mathbf{x}_T|M_2)p(M_1)}{B_{12}q_T(\mathbf{x}_T|M_2)p(M_1) + q_T(\mathbf{x}_T|M_2)p(M_2)} = \frac{\frac{B_{10}}{B_{20}}p(M_1)}{\frac{B_{10}}{B_{20}}p(M_1) + \frac{B_{20}}{B_{20}}p(M_2)} = \frac{B_{10}p(M_1)}{B_{10}p(M_1) + B_{20}p(M_2)}$. Since $2 \log B_{10} = BIC_0 - BIC_1 = -BIC_1$, then $B_{10} = \exp(-\frac{1}{2}BIC_1)$, the expression then becomes $p(M_1|\mathbf{x}_T) = \frac{\exp(-\frac{1}{2}BIC_1)p(M_1)}{\sum_{j=1}^2 \exp(-\frac{1}{2}BIC_j)p(M_j)}$.

Extending from 2 models to K models, $p(\mathbf{x}_T|M_k) \propto \exp(-\frac{1}{2}BIC_k)$ and (39) becomes

$$p(M_k|\mathbf{x}_T) = \frac{\exp(-\frac{1}{2}BIC_k)p(M_k)}{\sum_{j=1}^K \exp(-\frac{1}{2}BIC_j)p(M_j)}. \quad (40)$$

The expression in (40) uses the “full information” BIC derived in (34). In the framework of our GMM analysis, and following the discussion in Section 3.2 we modify (40)

¹⁷For example, the summation and implicit integrations in (39) below may be difficult to compute. Proposed solutions to this problem are novel Markov Chain Monte Carlo techniques such as the MC3 sampler first used by Madigan and York (1995) or averaging over a subset of models that are supported by the data such as the Occam’s window method of Madigan and Raftery (1994).

¹⁸In the last section we will address how our proposed methodology addresses some of the weaknesses of the BMA, and how it compares to the approaches of Brock and Durlauf (2001) and Doppelhofer, Miller, and Sala-i-Martin (2000).

¹⁹Note that $B_{00} = 1$ and $BIC_0 = 0$.

to incorporate the “limited information” LIBIC defined in (37)

$$p(M_k|\mathbf{x}_T) = \frac{\exp(-\frac{1}{2}LIBIC_k)p(M_k)}{\sum_{j=1}^K \exp(-\frac{1}{2}LIBIC_j)p(M_j)}. \quad (41)$$

Equation (41) defines the LIBMA estimator, an extension of the BMA in the case of a limited information likelihood. The LIBMA incorporates a dynamic panel estimator in the context of GMM and a Bayesian robustness check to explicitly account for model uncertainty in evaluating the results of a universe of models generated by a set of possible regressors.

V. STATISTICS FOR THE ROBUSTNESS ANALYSIS

This section summarizes the computational aspects and introduces the statistics on which we will base our robustness analysis.

Suppose we have n independent replications of a linear regression model with an intercept α , and k^* possible regressors grouped in a k^* -dimensional vector β . Denote by Z the corresponding $n \times k^*$ design matrix. We have $K = 2^{k^*}$ possible sampling models, depending on whether we include or exclude each of the regressors, so we have the space of all possible models $\mathcal{M} = \{M_j : j = 1, \dots, 2^{k^*}\}$. In order to deal with model uncertainty in the Bayesian framework, we need to define a prior distribution over the model space \mathcal{M} , namely, $p(M_j)$, where $\sum_{j=1}^{2^{k^*}} p(M_j) = 1$.

A model M_j with $0 \leq k_j \leq k^*$ regressors is defined by $y = \alpha + \beta_j X_j + \varepsilon$, where y is the vector of observations, X_j denotes the $n \times k_j$ matrix of the regressors included, and β_j is the vector of the relevant coefficients.

From (23) the posterior odds ratio for two models M_j, M_l is $B_{jl} = \frac{p(M_j|\mathbf{x}_T)}{p(M_l|\mathbf{x}_T)} = \frac{q_T(\mathbf{x}_T|M_j)}{q_T(\mathbf{x}_T|M_l)} \times \frac{p(M_j)}{p(M_l)}$. The first term on the right hand side, $\frac{q_T(\mathbf{x}_T|M_j)}{q_T(\mathbf{x}_T|M_l)}$ is the Bayes factor and can be approximated using (34). The second term, $\frac{p(M_j)}{p(M_l)}$ is the prior odds ratio. In the case where there is no preference for a specific model, $p(M_1) = p(M_2) = \dots = p(M_K) = \frac{1}{K}$ and the posterior odds ratio is equal to the Bayes factor. We *do not* assume equal inclusion probability for each model. Instead, following Doppelhofer, Miller, and Sala-i-Martin (2000) we represent a model M_j as a length k^* binary vector in which a one indicates that a variable is included in the model and a zero indicates that it is not. In addition, following Doppelhofer, Miller, and Sala-i-Martin (2000), we do not require the choice of (arbitrary) priors for all the parameters – instead, only one hyper-parameter is specified, the expected model size, \bar{k} .

Assuming that each variable has an equal inclusion probability, the prior probability for model M_j is

$$p(M_j) = \left(\frac{\bar{k}}{k^*}\right)^{k_j} \left(1 - \frac{\bar{k}}{k^*}\right)^{1-k_j}, \quad (42)$$

and the prior odds ratio is

$$\frac{p(M_j)}{p(M_l)} = \left(\frac{\bar{k}}{k^*}\right)^{k_j - k_l} \left(1 - \frac{\bar{k}}{k^*}\right)^{k_l - k_j}, \quad (43)$$

where k^* is the total number of regressors, \bar{k} is the researcher's prior about the number of regressors with non-zero coefficients, k_j is the number of included variables in model M_j , and $\frac{\bar{k}}{k^*}$ is the prior inclusion probability for each variable. Since \bar{k} is the only prior that arbitrarily specified in the simulations, robustness checks of the results can be ran by changing the value of this parameter.

If the set of possible regressions is small enough to allow exhaustive calculation, we can substitute (42) into (44) to calculate the posterior model probabilities (where the weights for different models are assigned based on posterior probabilities of each model—essentially normalizing the weight of any model by the sum of the weights of all possible $K = 2^{k^*}$ models) so that:

$$p(M_j|\mathbf{x}_T) = \frac{\exp(-\frac{1}{2}LIBIC_j)p(M_j)}{\sum_{l=1}^{2^{k^*}} \exp(-\frac{1}{2}LIBIC_l)p(M_l)}. \quad (44)$$

Next, we can use (44) to estimate the *posterior mean* and *posterior variance* as follows:

$$E(\theta_k|\mathbf{x}_T) = \sum_{j=1}^{2^{k^*}} p(M_j|\mathbf{x}_T) E(\theta_k|\mathbf{x}_T, M_j), \quad (45)$$

and

$$\begin{aligned} Var(\theta_k|\mathbf{x}_T) &= E[Var(\theta_k|\mathbf{x}_T, M_j)|\mathbf{x}_T] + Var[E(\theta_k|\mathbf{x}_T, M_j)|\mathbf{x}_T] \\ &= \sum_{j=1}^{2^{k^*}} p(M_j|\mathbf{x}_T) \{Var(\theta_k|\mathbf{x}_T, M_j) + E(\theta_k|\mathbf{x}_T, M_j)^2\} - E(\theta_k|\mathbf{x}_T)^2. \end{aligned} \quad (46)$$

Other statistics relevant to the study are the posterior mean and variance *conditional on inclusion*. First we calculate the *posterior inclusion probability*, which is the sum of all posterior probabilities of all the regressions including the specific variable (regressor). The posterior inclusion probability is a ranking measure to see how much the data favors the inclusion of a variable in the regression, and is calculated as

$$posterior\ inclusion\ probability = p(\theta_k \neq 0|\mathbf{x}_T) = \sum_{\theta_k \neq 0} p(M_j|\mathbf{x}_T). \quad (47)$$

If $p(\theta_k \neq 0|\mathbf{x}_T) > p(\theta_k \neq 0) = \frac{\bar{k}}{k^*}$ then the variable has a high marginal contribution to the goodness of fit of the regression model. Then, the posterior mean and variance conditional on inclusion are the ratios of the posterior mean and variance divided by the posterior inclusion probability, $\frac{E(\theta_k|\mathbf{x}_T)}{\sum_{\theta_k \neq 0} p(M_j|\mathbf{x}_T)}$, and $\frac{Var(\theta_k|\mathbf{x}_T)}{\sum_{\theta_k \neq 0} p(M_j|\mathbf{x}_T)}$, respectively.

Finally, we compute the *sign certainty probability*. This measures the probability that the coefficient is on the same side of zero as its mean (conditional on inclusion) and is calculated as

$$\begin{aligned} \text{sign certainty for } \theta_k &= p[\text{sgn}(\theta_k) = \text{sgn}E(\theta_k|D)|D, \theta_k \neq 0] & (48) \\ &= \sum_{j=1}^{2^{k^*}} p(M_j|\mathbf{x}_T) \{p[\text{sgn}(\theta_k) = \text{sgn}E(\theta_k|\mathbf{x}_T)|M_j, \mathbf{x}_T]\}. \end{aligned}$$

VI. CONCLUSION

This paper develops the theoretical background of the Limited Information Bayesian Model Averaging (LIBMA) approach and the computational aspects of the robustness analysis. The proposed methodology consists of a coherent Bayesian framework that addresses the problems of model uncertainty and restrictive assumptions of certain estimation procedures. The LIBMA technique has many potential applications including investigations of competing hypotheses, and parameter estimation that is robust to model specification.

As is typical in many areas of economic research, empirical work on investigating growth (and poverty) determinants is (i) prone to inconsistent estimates due to bias from omitted country-specific effects and failing to account for endogenous regressors; and (ii) particularly susceptible to model uncertainty arising from the combination of a complex web of relationships and the lack of clear theoretical guidance on the choice of regressors. The first practical application of the LIBMA by Ghura, Leite, and Tsangarides (2002) is a contribution to the ongoing growth and poverty debate that provides empirical evidence on the elasticity of the income of the poor with respect to average income and on the set of macroeconomic policies that directly influence poverty rates. Further, motivated by the existing empirical evidence on poverty reduction (and more broadly on human development), which strongly supports the primacy of the role of economic growth, a second research project attempts to explain the observed differences in standards of living across countries by identifying robust patterns of cross-country growth behavior, and examine convergence using the LIBMA approach.

References

- Andrews, D., and J. C. Monahan, 1992, "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, Vol. 60 (July), pp. 953-66.
- Brock, W., and S. Durlauf, 2001, "Growth Empirics and Reality," *World Bank Economic Review*, Vol. 15 (no. 2), pp. 229-72.
- Csiszar, I., 1975, "I-divergence Geometry of Probability Distributions and Minimization Problems," *Annals of Probability*, Vol. 3 (no. 1), pp. 146-58.
- Doppelhofer, G., R. I. Miller, and X. Sala-i-Martin, 2000, "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," Working Paper No. 7750 (Cambridge, Massachusetts: National Bureau of Economic Research).
- Ghura, D., C. Leite, and C. Tsangarides, 2002, "Is Growth Enough? Macroeconomic Policy and Poverty Reduction," IMF Working Paper No. 02/118 (Washington: International Monetary Fund).
- Golan, A., G. Judge, and D. Miller, 1996, *Maximum Entropy Econometrics: Robust Estimation with Limited Data* (Chichester: Wiley & Sons).
- Hansen, L. P., 1982, "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, Vol. 50 (July), pp. 1029-54.
- Hoeting, J. A., and others, 1999, "Bayesian Model Averaging: A Tutorial," *Statistical Science*, Vol. 14 (No. 4), pp. 382-417.
- Inoue, A., 2001, "A Bayesian Method of Moments in Large Samples," Working Paper (Raleigh: North Carolina State University).
- Kass, R. and A. Raftery, 1995, "Bayes Factors," *Journal of the American Statistical Association*, Vol. 90 (no. 430), pp. 773-95.
- Kim, J. Y., 2000, "The Generalized Method of Moments in the Bayesian Framework and a Model and Moment Selection Criterion," Working Paper (Albany: State University of New York).
- _____, 2001, "Bayesian Limited Information Analysis in the GMM Framework," Working Paper (Albany: State University of New York).
- _____, 2002, "Limited Information Likelihood and Bayesian Analysis," *Journal of Econometrics*, Vol. 107 (March), pp. 175-93.

- Leamer, E., 1978, *Specification Searches: Ad Hoc Inference with Non-experimental Data*, (New York: Wiley & Sons).
- Madigan, D., and A. Raftery, 1994, "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of American Statistical Association*, Vol. 63, pp. 1535-46.
- Newey, W., and K. West, 1987, "Hypothesis Testing with Efficient Method of Moments Estimation," *International Economic Review*, Vol. 28 (no. 3), pp. 777-87.
- Raftery, A. E., 1994, "Bayesian Model Selection in Social Research," in *Sociological Methodology*, ed. by Peter V. Marsden, (Cambridge, Massachusetts: Blackwells), pp. 111-96.
- _____, 1996, "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," *Biometrika*, Vol. 83, pp. 251-66.
- Tierney L., and J. B. Kadan, 1986, "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of American Statistical Association*, Vol. 81, pp. 82-86.
- White, H., 1980, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, Vol. 48 (May), pp. 817-38.
- _____, 1982, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, Vol. 50 (January), pp. 1-25.
- Zellner, A., 1996, "Bayesian Method of Moments/Instrumental Variable (BMOM/IV) Analysis of Mean and Regression Model," in *Modeling and Prediction Honoring Seymour Geisser*, ed. by J.C. Lee, W.C. Johnson, and A. Zellner (New York: Springer), pp. 61-74.
- _____, 1997, "The Bayesian Method of Moments (BMOM): Theory and Application," in *Advances in Econometrics Vol. 12*, ed. by T. Fomby, R. C. Hill (Greenwich, Connecticut: JAI press), pp. 85-105.