European Economic Review 81 (2016) 86–102

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

Growth econometrics for agnostics and true believers ☆

James Rockey [a], Jonathan Temple [b,*]

[a] University of Leicester, United Kingdom
[b] University of Bristol, United Kingdom

ARTICLE INFO

Available online 25 June 2015

JEL classification:
C51
O40
O47

Keywords:
Bayesian Model Averaging
Growth regressions
Growth econometrics

ABSTRACT

The issue of model uncertainty is central to the empirical study of economic growth. Many recent papers use Bayesian Model Averaging to address model uncertainty, but (Ciccone and Jarociński, 2010) have questioned the approach on theoretical and empirical grounds. They argue that a standard 'agnostic' approach is too sensitive to small changes in the dependent variable, such as those associated with different vintages of the Penn World Table (PWT). This paper revisits their theoretical arguments and empirical illustration, drawing on more recent vintages of the PWT, and introducing an approach that limits the degree of agnosticism.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In principle, an interesting and perhaps salutary history could be written of the details of statistical practice, and commentary on that practice. The topic of data mining, in the prejudicial sense, would probably loom large. When Leamer (1983) drew attention to the weaknesses of empirical work in economics, he titled his paper 'Let's take the con out of econometrics'. His particular targets were the difficulties associated with observational data, the problems raised for classical inference by data mining and specification searches, and the dependence of results on questionable assumptions. In his view, applied econometricians rarely did enough to examine or communicate these assumptions, or the sensitivity of the findings to reasonable alternatives. He wrote 'This is a sad and decidedly unscientific state of affairs we find ourselves in. Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously' (Leamer, 1983, p. 37).

How much has changed? More than thirty years later, much published work remains hard to interpret, because readers are aware that the reported models may be the outcome of data mining and specification searches. The reported standard errors and confidence intervals are rarely adjusted for the effects of model selection, and it is far from clear whether they should be taken seriously, or why alternative models have been ruled out. Leamer (1978) had outlined the principles of a solution, namely to average across models on Bayesian principles, but for a long time this appeared computationally intractable.

The development of practical methods for Bayesian Model Averaging (BMA), initiated in particular by the work of Raftery (1995) and Raftery et al. (1997), represented a considerable advance. The approach assigns a prior probability to each of a set

☆ This paper contributes to the literature on Bayesian Model Averaging, an area where the late Eduardo Ley made a series of influential contributions. Other debts are due to the editors, two anonymous referees, Paddy Carter, Antonio Ciccone and Adeel Malik for comments and suggestions that have helped to improve the paper. We are responsible for its remaining shortcomings.

* Corresponding author.
E-mail address: jon.temple@bristol.ac.uk (J. Temple).

http://dx.doi.org/10.1016/j.euroecorev.2015.05.010
0014-2921/© 2015 Elsevier B.V. All rights reserved.

of models, updates these prior probabilities in the light of the data, and then averages across the models using the posterior model probabilities. To assess the robustness of the evidence that a variable is relevant, the posterior model probabilities can be summed across all models that include that variable, yielding the variable's 'posterior inclusion probability'. This approach provides researchers with a rich array of information about the extent of model uncertainty and its implications for the substantive conclusions. The early applications in Brock and Durlauf (2001) and Fernández et al. (2001a) showed the potential of BMA for the study of growth, and it has become a popular approach, aided by computer software that makes it easy to implement.

In an important paper, Ciccone and Jarociński (2010) take a more sceptical view. Their arguments are both theoretical and empirical. They point out that the most common version of BMA uses priors which make posterior inclusion probabilities too sensitive to small changes across models in the sum of squared errors. This can happen even if these changes are unsystematic, perhaps arising from noise in the measurement of the dependent variable. They demonstrate this effect in action by combining the data set of Sala-i-Martin et al. (2004) with three different vintages of the output data from the Penn World Table (PWT). By computing the growth rate and initial GDP per capita using alternative vintages of the data, they show that posterior inclusion probabilities become surprisingly unstable. They write:

> Overall, our findings suggest that margins of error in the available income data are too large for empirical analysis that is agnostic about model specification. It seems doubtful that the available international income data will tell an agnostic about the determinants of economic growth. (Ciccone and Jarociński, 2010, p. 244).

Is this conclusion too strong? The paper by Ciccone and Jarociński, in its combination of theoretical arguments and a striking case study, is an important critique of BMA applied to cross-country data. It seems to have modified perceptions of the approach. Drawing on their work, Easterly's remarks, prompted by Banerjee (2009), may be representative:

> In the parade of ignorance about growth, 145 variables were 'significant' in growth regressions. Since a cross-section regression has about 100 observations, Banerjee rightly notes that growth researchers work with negative degrees of freedom. Some attempted to reduce the set to a much smaller number of robust variables using Bayesian Model Averaging, which raised hopes briefly. But this approach gave completely different 'robust' variables for different equally plausible samples. More than Banerjee acknowledges, macroeconomists have earned their ignorance the hard way. Easterly (2009, p. 228), notes omitted.

In responding to the work of Ciccone and Jarociński (henceforth CJ), our paper makes a number of arguments. First, for reasons we explain below, the theoretical argument that BMA necessarily relies on a small set of models can be qualified. Second, we point out that posterior inclusion probabilities in CJ vary partly because initial GDP per capita and regional dummies move in and out of the models. For a variety of reasons, these movements risk exaggerating the underlying instability across vintages of the data. When we address this problem, we indeed find that the instability of the posterior inclusion probabilities is reduced. This remains the case when we add further PWT vintages that were not available to CJ, namely PWT 6.3, 7.0, 7.1 and 8.0. Hence, we recommend that future work seeking robust growth determinants should include initial GDP per capita and regional dummies in each model considered; this helps to guard against the 'false positives' that can otherwise emerge.

As this may indicate, our paper is best seen as qualifying and extending CJ's arguments rather than rebutting them. CJ suggest that, given the weaknesses of the data, agnostic approaches using BMA cannot reliably identify growth determinants, and often point towards variables whose importance turns out to be fragile. Our contribution is to show that restricting the set of candidate models can limit the instability of BMA and the extent of potential false positives. This supports a broader implication of CJ, that one can have too much agnosticism. Also like them, we can see a case for modifying standard BMA approaches to increase the robustness of the results. Future research is likely to draw heavily on improved methods such as those advocated in Feldkircher and Zeugner (2009, 2012) and Ley and Steel (2009, 2012). In the meantime, given the large number of empirical studies that have already been published using BIC approximations or closely-related methods, it remains worth investigating whether CJ's points are widely applicable. Is Easterly right to conclude that BMA briefly 'raised hopes' which have now been dashed?

The remainder of the paper has the following structure. Section 2 sketches the background to our paper in more detail. Section 3 discusses the theoretical arguments made in CJ. The heart of the paper, Section 4, revisits their study of the Sala-i-Martin et al. (2004) data using the same vintages of the Penn World Table, and then extends it using later vintages. Section 5 discusses some broader issues raised by the use of BMA to analyse growth data. Finally, Section 6 concludes with some practical recommendations. In effect, we argue that studies of economic growth should either assign a high prior inclusion probability to initial GDP per capita and region fixed effects, or go even further, and require these variables to be included in all the models considered.

2. Background

As is well known, outside rare special cases, basing inference on a single model is problematic. Conventional inference will underestimate the extent of uncertainty about the parameters, and confidence intervals will be consistently too narrow (see, for example, Leamer, 1974; Gelman and Rubin, 1995, p. 170). But the problem of model uncertainty does not end there.

The regressions reported in applied research in economics are typically a subset of those that have been estimated. Researchers might use an 'undirected' model search in which approaches such as stepwise regression, or model selection criteria, are used to identify a small set of models, without actively pursuing a particular finding. In economics, however, it seems likely that a significant proportion of published results have arisen from a 'directed' model search. Results are selectively reported in whichever way best serves the objectives of the researcher, which (almost inevitably) extend beyond the generation of reliable evidence.[1] Inference is then carried out as if the chosen model generated the data and was the only one considered. The influential statistician Leo Breiman called this type of practice a 'quiet scandal' in statistics (Breiman, 1992, p. 738).

When the data have been used to select models as well as estimate them, inference should take this into account. It is likely that the best approach will be to treat model selection, estimation and inference as a combined effort, as emphasized in Magnus and de Luca (2014). Moreover, the most attractive approaches will be those which acknowledge the researcher's inevitable uncertainty about which model is the best approximation to the data generating process. In turn, this seems to point to using information from multiple models, often using weighted averages across a variety of models. An analogy from Claeskens and Hjort (2008, p. 192–193) helps to justify this. Faced with a set of statisticians who provide conflicting evidence on a question of interest, should one try to identify the best statistician and listen only to them, or should one accept that identifying the best will be difficult and subject to error, and instead decide to aggregate their evidence?[2]

At least in outline, the problem of model uncertainty has been appreciated for a long time in the growth literature. The decisive paper was Levine and Renelt (1992), which applied a version of extreme bounds analysis (EBA) to cross-country growth regressions. Their paper is one of the most influential in the field, and has been cited in around 1300 published articles and more than a hundred books.[3] The paper remains the most famous implementation of an EBA, but papers using this approach continue to appear, such as Gassebner et al. (2013).

The underlying idea, due to Leamer (1983) and Leamer and Leonard (1983), is to study whether regression coefficients remain significant when different combinations of control variables are tried, where these are drawn from a large set of candidates. The alternative sets of control variables lead to lower and upper bounds on the coefficient of interest, based on the extremes of confidence intervals; a natural question is whether these bounds have the same sign, and therefore exclude zero.

The application to growth in Levine and Renelt (1992) indicates that zero can rarely be excluded. This has often been summarized as 'nothing is robust', but in some respects this goes too far. A general problem with extreme bounds analysis is that variables may be rendered insignificant within specifications that are themselves flawed or weak in explanatory power. In principle, if enough candidate variables are considered, any empirical finding can be overturned, whatever its underlying validity. A related, and fundamental, problem is that it is hard to provide statistical foundations for this approach.

Some of the problems can be illustrated by considering the Levine and Renelt results in more detail. More than a third of the upper and lower bounds are generated by regressions that include the growth rate of domestic credit as a control variable. This variable is likely to be strongly endogenous, and positively correlated with output growth. It is not altogether surprising that its inclusion sometimes renders other variables insignificant, and this example illustrates the pitfalls of allowing 'too much' competition from other models.[4] It also shows that agnosticism should not preclude careful attention to the set of candidate models.

A deeper objection to EBA is that its emphasis on significance at conventional levels is conceptually problematic, especially for policy decisions. As Brock and Durlauf (2001), Brock et al. (2003) and Cohen-Cole et al. (2012) have argued, the conventional dichotomy between significant and insignificant variables makes unexamined, and frequently implausible, assumptions about the relative costs of Type I and Type II errors. Ideally, a decision-maker should think in terms of a loss function and attach probability distributions to the parameters associated with particular variables. That approach is a natural counterpart or sequel to Bayesian Model Averaging.

Versions of BMA have been implemented for growth data by Brock and Durlauf (2001), Danquah et al. (2014), Eicher et al. (2007), Fernández et al. (2001a), Masanjala and Papageorgiou (2008), Sala-i-Martin et al. (2004), Sirimaneetham and Temple (2009) and many other authors. More recent applications have gone beyond national growth determinants, to include regional growth (Crespo Cuaresma et al., 2014), productivity forecasting (Bartelsman and Wolf, 2014), reform indicators (Kraay and Tawara, 2013), business cycle models (Strachan and van Dijk, 2013), exchange rate forecasting (Wright, 2008; Garratt and Lee, 2010), the effects of trade agreements (Eicher et al., 2012), international migration (Mitchell et al., 2011), environmental economics (Begun and Eicher, 2008), financial development (Huang, 2011), output volatility (Malik and Temple, 2009), the country-specific severity of the global financial crisis (Feldkircher, 2014), the effects of the crisis on exchange rates (Feldkircher et al., 2014), the deterrent effect of capital punishment (Durlauf et al., 2012), and corporate default rates (Gonzlez-Aguado and Moral-Benito, 2013).

[1] If this claim appears sweeping or a little too cynical then see, for example, the odd distribution of published *p*-values documented in Brodeur et al. (2015). The paper by Fanelli (2010) finds that published papers in the social sciences are more likely to report support for hypotheses than papers in many of the sciences, although in many cases, the differences across disciplines are modest.

[2] There is a sense in which averaging across many different models is linked to the 'wisdom of crowds' discussed in Surowiecki (2004).

[3] Figures from Web of Science, as of April 2015. Citations tracked by Google Scholar exceed 6400.

[4] This excess competition problem had been recognized in the earlier literature on EBA; see Temple (2000) for discussion and references. The problem is not simply competition from weak models, but also competition from models that have 'too much' explanatory power, because of the inclusion of an endogenous variable.

Table 1
Malik and Temple (2009) results.[a]

Model	Dimension	R^2	Adjusted R^2	Component	Weight
1	7	0.5765	0.5287	2.720E+11	1
2	8	0.5963	0.5433	8.183E+11	3.01
3	8	0.5959	0.5429	7.936E+11	2.92
4	7	0.5685	0.5198	1.413E+11	0.52
5	6	0.5395	0.4956	2.529E+10	0.09
6	9	0.6142	0.5564	2.266E+12	8.33
7	7	0.5631	0.5138	9.151E+10	0.34
8	8	0.5867	0.5324	3.584E+11	1.32
9	9	0.6107	0.5523	1.642E+12	6.04
10	9	0.6107	0.5523	1.642E+12	6.04

[a] This table shows, for the top ten models in Malik and Temple (2009), their dimension, the R^2, the adjusted R^2, and the components/weights used in building up the posterior inclusion probability.

Researchers have also extended the approach to allow more flexibility in the priors (Ley and Steel, 2009, 2012), to adapt the method for panels (Moral-Benito, 2012), to analyze regressors with interdependent roles (Doppelhofer and Weeks, 2009; Ley and Steel, 2007), and to address endogenous regressors (Koop et al., 2012), variable transformations and parameter heterogeneity (Eicher et al., 2007; Gottardo and Raftery, 2009; Salimans, 2012) and outliers and fat-tailed error distributions (Doppelhofer and Weeks, 2011; Gottardo and Raftery, 2009). There are also parallel literatures which develop or compare alternative ways to address model uncertainty: see, for example, Amini and Parmeter (2012), Deckers and Hanck (2014), Hansen (2007, 2014), Hendry and Krolzig (2004), Magnus et al. (2010), and Magnus and Wang (2014).

Our own paper is mainly about methods, and in particular, methods for identifying growth determinants from cross-section regressions. The questions of interest are whether, and how, the literature on this topic should respond to the instability identified by CJ. The problem is framed as one of uncovering a set of variables that have explanatory power for growth, in a context of substantial model uncertainty. On some occasions, there is a different question of interest, namely whether the effect of a specific variable is robust to different possible combinations of control variables (potential 'confounders'). Crainiceanu et al. (2008) call this related, but distinct, problem that of 'adjustment uncertainty'; it has obvious links to the EBA approach. Separately, Jensen and Würtz (2012) develop a method for estimating the effect of a variable when there are more candidate explanatory variables than observations.

3. Is BMA unstable?

In this section, we discuss the main theoretical argument of CJ. They propose that instability is likely in the standard version of Bayesian Model Averaging, because the posterior inclusion probabilities are highly sensitive to the sum of squared errors (SSE) generated by each model. Small changes to the dependent variable can then have large effects. They use this argument to explain why, in their application, posterior inclusion probabilities are sensitive to the vintage of the PWT used to construct the dependent variable and the measure of initial GDP per capita. We think these arguments are important, but introduce some additional considerations that qualify them a little.

One interpretation of CJ is that the posterior model probabilities will be negligible for all but the first-ranked model in a given size class. Our main point is that, when a wide range of models are considered, it is likely that the second-ranked model within a given size class will often have explanatory power that is close to the first-ranked model, and this will be reflected in the posterior model probabilities. The extent to which this is true will depend on the data set, but sometimes it will weaken the dominance of a single model in a given size class, as we illustrate below.

The theoretical argument in CJ starts from the observation that, for a fixed model size, the posterior inclusion probability (PIP) for a given variable z depends on sums of the form

$$PIP(z) \propto \sum_{j \in S_z} SSE_j^{-N/2}$$

where S_z denotes the set of models that contain the variable z, and N is the sample size. But when N is at least moderately large, this sum will be dominated by the best-fitting model, since $\sum_{j \in S_z} SSE_j^{-N/2} \approx \max SSE_j^{-N/2}$. This implies instability: small perturbations to the data will modify the sum of squared errors generated by each model, and may also change the identity of the best-fitting model. CJ argue that this explains why, in their application of BMA to economic growth, posterior inclusion probabilities vary widely across different vintages of the PWT. Feldkircher and Zeugner (2009) refer to the dominance of the best-fitting model as the 'supermodel effect'.

To examine this in more detail, we take another example from the cross-country literature. Malik and Temple (2009) study output volatility in a sample of 70 countries using a standard version of Bayesian Model Averaging. A BIC approximation is used to compare models, and the prior over models is uniform: before analysing the data, all models are considered equally likely, corresponding to a prior inclusion probability of 0.5 for each candidate variable. This approach is computationally simple and is that introduced by Raftery et al. (1997).

Table 2
Posterior inclusion probabilities.[a]

Variable	PWT 6.2	PWT 6.1	PWT 6.0
GDP per capita 1960	**1.00**	**1.00**	**0.69**
Primary schooling 1960	**1.00**	**0.99**	**0.79**
Fertility 1960s	**0.91**	**0.12**	0.03
Africa dummy	**0.86**	**0.18**	**0.15**
Fraction Confucius	**0.83**	**0.12**	**0.20**
Fraction Muslim	**0.40**	**0.19**	**0.11**
Latin American dummy	**0.35**	0.07	**0.14**
East Asian dummy	**0.33**	**0.78**	**0.83**
Fraction Buddhist	**0.28**	**0.11**	**0.11**
Primary exports 1970	**0.27**	**0.21**	0.05

[a] This table shows the first ten rows of PIPs from CJ (2010), their Table 2.

In Table 1, we show some key data for the top ten models in the Malik and Temple study—those with the highest posterior model probabilities, shown in their Table 2. Under CJ's argument that one model will heavily dominate for a given model size, the second-ranked model in that size class should have a substantially lower posterior model probability. Hence, it seems quite likely that the top ten models should be of different sizes. In fact, Table 1 shows that there are three models of dimension 9, three of dimension 8, three of dimension 7 and one of dimension 6.

The likely explanation can also be found in Table 1. When there are many candidate explanatory variables, there are many possible models—in this case $2^{23} \approx 8.39$ million candidate models. This suggests that a large number of models of a given size will have a similar *SSE* or, equivalently here, a similar R^2. This pattern is clear in Table 1, where models of a given dimension vary relatively little in the R^2 (compare models 2 and 3, or 9 and 10, for example). Whether or not this happens in other applications will depend on the data, but it can be seen as a natural consequence of having many models to choose from. For example, consider a case where one of the candidate variables is especially important for generating a high R^2. At first glance, it might seem that one of the models will have by far by the best fit — except that, when there are many candidate predictors, there could well be many ways of constructing a set of control variables that will lead to approximately the same R^2. Much the same argument implies that adding a particularly important variable to a data set will still lead the posterior mass to be distributed across a range of models. These considerations suggest that, at least for some data sets, the supermodel effect will be less extreme than the theoretical argument in Ciccone and Jarociński implies.

Another factor promoting stability is that the posterior inclusion probability for a given variable will draw on evidence from models of different sizes. This may not be sufficient, however, to allay concerns. If we again take the example of Malik and Temple (2009), their top ten models have a combined posterior probability of almost 40%. This seems implausibly high, given the vast number of possible models. It suggests that one standard approach to BMA favours a small set of models too readily, as CJ warned. Later in the paper, we will discuss how the recent literature has addressed this problem. For now, it is worth noting that conventional methods for reporting empirical results — methods that do not formally address model uncertainty — place even more weight on a small number of models; and these models often emerge from an unsystematic and perhaps 'prejudiced' search (Leamer, 1974). Hence, a tendency for this version of BMA to downweigh too many models is a reason for doing BMA differently, not for abandoning BMA altogether.

We explore the issues in more detail by taking CJ's argument at face value, while considering a BIC-based prior for simplicity. To do this, we ignore all models other than the dominant model (the best-fitting) within each size class. Define P_q as the prior probability attached to a model of size $q = 1, 2, \ldots, Q$. With a slight abuse of notation, define an indicator function $\Lambda_q(z)$ which is equal to one if the dominant model of size q contains the variable z, and is otherwise zero. For normal linear models, it is easy to show that the posterior inclusion probability for a particular variable z is proportional to the following quantity:

$$PIP(z) \propto \sum_{q=1}^{Q} (1-R_q^2)^{-N/2} \left(\frac{1}{\sqrt{N}}\right)^{q-1} P_q \Lambda_q(z)$$

There are several points to note, which are useful in thinking about CJ's results.

First, if the data for the dependent variable are modified, the posterior inclusion probability will change for two reasons: (1) changes in the R^2 of the various dominant models, and (2) changes in the Λs — that is, in whether the variable z does, or does not, enter into the new dominant model for each size class. This suggests that instability in the posterior inclusion probability is more likely to arise for variables of 'intermediate' robustness, rather than those excluded from most models (most of the Λs equal zero) or nearly always included (most of the Λs equal one). Again, this supports CJ's concerns, by indicating how instability can arise. But it is not a reason for abandoning BMA; if there are variables whose explanatory power is sensitive to the specification, this is exactly where conventional *ad hoc* methods for selecting and reporting models are most likely to go astray.

Second, as we move from one size class to the next in the summation, the increment to the posterior inclusion probability depends on the R^2 for the model that has an extra parameter, but with weights that decline geometrically. The

term in $1/\sqrt{N}$ represents the penalty for loss of parsimony. The penalty is substantial and, for samples of the size considered here, much larger than the penalty embedded in the adjusted R^2 (see the Appendix). The nature of the penalty in the BIC is conventionally justified by the ability of the BIC to approximate (twice the log of) Bayes factors in certain settings, building on the original argument of Schwarz (1978). But alternative penalties can be justified on alternative criteria, and this suggests that a more flexible approach might be beneficial, including approaches which are more agnostic about the appropriate penalty for loss of parsimony (Ley and Steel, 2009, 2012).

Third, but less fundamentally for the current paper, the expression casts some light on the role of the prior over models. Sala-i-Martin et al. (2004) argued that the uniform model prior, in which all models are initially judged equally likely, risks favouring models of a certain size: the prior mass over the model space will be heavily concentrated around a particular size. But when the chosen priors for the slope parameters give the most weight to the best-performing model within each size class, many of the other posterior model probabilities within that size class will effectively be rounded down to zero. Hence, the posterior inclusion probabilities will not be dominated by models of a certain size even when the prior model probabilities (the P_q in our notation) are the same across size classes.

Finally, note that using prior model probabilities that vary with model size will complicate the relationship between the posterior model probabilities and the BIC, since it effectively alters the penalties for model complexity.[5] When prior probabilities differ across models, this implies the P_q terms in the summation above will vary. This will make the posterior inclusion probabilities more sensitive to the results of some model class sizes (those with high P_q) and less to others. It is possible that this could make the results more unstable, rather than less. Drawing on evidence from out-of-sample predictive performance and simulations, Eicher et al. (2010) argue that the uniform model prior works well. An alternative view, associated with Feldkircher and Zeugner (2009) and Ley and Steel (2009, 2012), is that it makes sense to use more flexible priors in which the data are allowed to influence the penalty imposed on complex models, through the prior inclusion probabilities (Ley and Steel, 2009) or the priors over slope coefficients (Feldkircher and Zeugner, 2009; Ley and Steel, 2012). More flexible priors could help to limit the instability identified by CJ, and we return to this issue later in the paper.

4. Results

Thus far, we have sketched some reasons why the theoretical arguments in Ciccone and Jarociński (2010) are not conclusive, based on some additional considerations. But this raises an obvious question. If Bayesian Model Averaging is likely to be more robust than they suggest, how can we explain the results in their empirical application? Why do CJ find that growth regressions estimated using alternative vintages of the Penn World Table lead to very different posterior inclusion probabilities?

Part of the explanation is that when CJ use different vintages, together with the largest possible sample, the composition of the sample varies. This is examined by Feldkircher and Zeugner (2012), who find that changes in the sample play an important role, not least because the countries which move in and out of the sample tend to be African ones that are especially likely to be outliers. Nevertheless, even after restricting attention to a fixed sample, some instability remains. Feldkircher and Zeugner recommend the use of alternative priors, to limit the importance of the supermodel effect. In particular, one aim of such priors, as in Ley and Steel (2009, 2012), is to increase robustness to prior assumptions, which in turn may limit the instability that CJ identified.

The recent literature makes a strong case for using more sophisticated priors. Nevertheless, given that simpler priors continue to have some support (Eicher et al., 2010), and have been widely used in the empirical literature, it remains interesting to examine CJ's results in more detail. In the remainder of the paper we will argue that, even with inflexible priors, restricting the set of candidate models can greatly increase the stability of BMA results.

In particular, we think there is a strong case for requiring the candidate models to include initial GDP per capita and regional dummies. To see this, consider Table 2, which lists the first ten sets of posterior inclusion probabilities listed in Table 2 of CJ's paper. An interesting feature of this table is the posterior inclusion probability for the logarithm of initial (1960) GDP per capita. This has a posterior inclusion probability of 1.00 for PWT 6.1 and 6.2, but only 0.69 in PWT 6.0. In other words, when a standard version of BMA is applied to the PWT 6.0 data, more than 30% of the posterior mass is accounted for by models that exclude initial GDP per capita. Since initial GDP per capita is likely to be quite strongly correlated with some of the candidate explanatory variables, this could be a source of instability in posterior inclusion probabilities.

For a study of growth determinants, averaging across models which sometimes exclude initial GDP per capita, and sometimes do not, is conceptually problematic. This is because the slope parameters have different economic interpretations, depending on whether or not initial GDP per capita is included. Averaging the parameters becomes problematic, given that the economic interpretation of the parameters varies across models.[6]

To give a specific example, consider the empirical implications of the Solow model, as developed in Mankiw et al. (1992). The model predicts that growth, measured over long time periods, will be uncorrelated with investment rates: this is the

[5] See George and Foster (2000) for analysis of how the priors over models, and the priors over slope coefficients, sometimes combine so that ranking by posterior model probabilities corresponds to ranking by various formal model criteria.

[6] In a related context, Hlouskova and Wagner (2013) have also noticed the importance of including initial GDP per capita in all models.

famous result that the long-run growth rate is independent of the investment rate. But when initial GDP per capita is included on the right-hand-side, the other right-hand-side variables do not have to be growth determinants: they can instead be variables which influence the height of the steady-state growth path. This set of variables could include, among others, the investment rate. A direct consequence is that, on standard theoretical grounds, the relevant set of predictors will depend on whether or not the regression controls for initial GDP per capita.

A more technical way to support this claim is to think about the underlying priors. Conceptually, the prior that a variable has a long-run growth effect is distinct from the prior that it influences the height of the steady-state growth path. These differences in priors should presumably be related to the strength of the researcher's belief that endogenous growth models (in which long-run growth effects are feasible) are likely to describe the data better than neoclassical growth models in the traditions of Ramsey and Solow (in which long-run growth effects are absent). If a model averaging exercise allows GDP per capita to move in and out of the candidate models, the two possibilities are conflated in ways that make the assumed priors hard to justify.

A further issue is that a uniform model prior may be implausible in this context. That prior will assign a prior probability of 0.5 to the inclusion of initial GDP per capita. In effect, endogenous growth models with long-run growth effects, and neoclassical growth models without long-run growth effects, are treated as equally likely before the data are analyzed. It is not clear that this prior reflects the current state of knowledge. Since the early work of Marris (1982), Kormendi and Meguire (1985), Baumol (1986), Barro (1991), Barro and Sala-i-Martin (1992) and Mankiw et al. (1992), empirical work on growth has routinely included a role for initial GDP per capita, and found it to be statistically significant. Under this approach, the explanatory variables are usually interpreted as giving rise to level effects—that is, they influence the height of the steady-state growth path—rather than growth effects. Moreover, it is known that, in theoretical models, the existence of long-run growth effects typically relies on knife-edge parameter assumptions; hence, level effects seem more plausible than growth effects.[7] This paper proposes that, in order to ensure parameters can be given the same economic interpretation across models, a model averaging exercise for growth should include initial GDP per capita in *all* the models considered. This can be justified partly on theoretical grounds, and partly by a large body of empirical work which finds initial GDP per capita to play a role, at least once steady-state determinants are included.[8]

Further examination of Table 2 makes clear another reason for instability of the posterior inclusion probabilities in CJ's study. Many of the variables listed are dummies for regions or dominant religion. There are several reasons why these variables could lead to instability in posterior inclusion probabilities. The most mechanical, although perhaps not the most important, is that different combinations of region dummies can give rise to an artificial form of model uncertainty. First consider the extreme case where a set of R region fixed effects is complete, in that each country is allocated to a region (as when zero/one region dummies sum to one for each country). Not all the region fixed effects can be included, given the presence of an intercept. But there are several (in fact, R) statistically equivalent ways of including $R-1$ region dummies, depending on which is the omitted region. Depending on the software package, models that are statistically equivalent may be treated as different models. This will give rise to an artificial form of model instability; for example, the posterior inclusion probabilities of the individual region dummies are likely to be unstable.[9] A more general version of this argument applies when several region dummies are included in the candidate explanatory variables, but not a complete set. They may still substitute for each other in models that are otherwise similar. These are particular instances of the jointness effects analysed in Doppelhofer and Weeks (2009) and Ley and Steel (2007).

More importantly, this argument can be taken further, and extended to other variables. Part of the conventional wisdom in the growth literature is that regions matter, in the following sense: two countries drawn from the same region will often look far more alike than two countries drawn at random from a global sample. This perception helps to explain why so much historical and political research tends to group countries on regional lines. For example, there is a long tradition of seeing sub-Saharan African countries as culturally, socially and economically distinct from the countries of the Middle East and North Africa. But within-region homogeneity has a practical consequence for model uncertainty: even some of the continuous explanatory variables, other than the region fixed effects, will exhibit more variation across regions than within them. Alternative variables will then move in and out of growth models, depending on the combinations that happen to be the best proxy for broader regional characteristics in that particular model. This will generate instability in the posterior inclusion probabilities, unless the models always include region fixed effects.

This argument may seem speculative, but consider the following exercise. We take the candidate growth determinants considered by CJ and regress each of these candidates on the four regional dummies in the data set: these are dummies for sub-Saharan Africa, East Asia, Europe, and Latin America. (This approach will be conservative, in that we could instead have

[7] These level effects may sometimes be large, and driven by the same mechanisms studied in endogenous growth models. Temple (2003) discusses these issues in more detail.

[8] This raises a deeper question: perhaps the parameters of interest in a model averaging exercise should be the long-run effects, defined as the slope coefficients divided by the absolute size of the coefficient on the logarithm of initial GDP per capita. Developing methods which define the priors over these long-run effects would be an interesting area for further work.

[9] Software packages for BMA, including the original 'bicreg' software associated with Raftery et al. (1997) and the BMS software of Feldkircher and Zeugner, vary in their treatment of subsets of variables when combinations are precisely collinear; and proceed as normal if collinearity is not exact, even if it is close to being present.

Table 3
Proportion of variation in each variable explained by four region dummies.[a]

Variable	R^2
Fraction population over 65	0.80
Absolute latitude	0.69
Fraction population less than 15	0.68
Life expectancy in 1960	0.67
Malaria prevalence in 1960s	0.67
Fertility in 1960s	0.62
Political rights	0.62
Fraction of tropical area	0.62
Fraction population in tropics	0.58
Years open 1950–1994	0.51
Primary schooling in 1960	0.50
Spanish colony	0.49
Fraction muslim	0.48
Higher education 1960	0.46
Fraction Catholic	0.45
Air distance to big cities	0.44
Civil liberties	0.43
Real exchange rate distortions	0.43
Ethnolinguistic fractionalization	0.43
Timing of independence	0.42

[a] This table reports the R^2 of the regression of each candidate variable on four region dummies, for Sub-Saharan Africa, Latin America, East Asia, and Europe.

considered a complete set of region fixed effects.) We are interested in the R^2 of each of these regressions: how much of the variation in a candidate growth determinant arises from variation across regions? The results are shown in Table 3, restricting attention to the twenty candidates with the highest R^2s. There are eight variables where the four region dummies account for more than 60% of the variation, and the R^2 is higher than 0.40 for all twenty variables in the table. Moreover, the twenty variables include primary schooling in 1960, the fraction Muslim, fertility in the 1960s, and primary exports in 1970, all of which are listed in our earlier Table 2. Hence, the instability of posterior inclusion probabilities shown in Table 2 may arise partly through the suggested mechanism. Given that some candidate growth determinants vary more across regions than within them, different combinations of these variables and region dummies will substitute for one another across models.

There is a simple solution to this problem, which is to require the candidate models to control for region effects. This helps to guard against 'false positives' that can otherwise emerge when a candidate growth determinant varies mainly across regions. Put differently, if a researcher wants to use cross-section regressions in preference to panel data models, region effects are a convenient (but only partial) substitute for the rigour of allowing a full set of country-specific effects. Temple (1998) argued that region fixed effects should be used to proxy for regional differences in an unobservable variable, initial efficiency, which in principle should be included in conditional convergence regressions (see Mankiw et al., 1992; Islam, 1995).

In a panel data setting, fixed effects are typically included because a researcher is concerned about omitted variables, including omitted variables that may be difficult to measure or whose importance has not been appreciated by the researcher. Some of this logic can be extended to a model averaging exercise. The set of candidate variables included in the analysis is likely to be incomplete. In the context of economic growth, a researcher can never rule out the possibility that a relevant variable has been omitted from the analysis. As Brock and Durlauf (2001) emphasize, growth theories are open-ended, in the sense that a role for one variable need not preclude the existence of additional forces. If we take the view that at least some of the omitted variables are likely to vary substantially across regions, and less so within them, then forcing the models to include region effects could help to address this point. We conjecture that it will improve the performance of model averaging methods when the set of candidate variables is incomplete.[10]

There are two main opposing arguments. First, when a researcher constructs regional groupings, this may draw on background knowledge of how different regions have grown, which could bias the results. From a statistical point of view, including a dummy based on current membership of the OECD would be questionable, but defining regions on conventional geographic grounds is more defensible. Second, one might be concerned that including region dummies means that the growth models speak to a different set of questions, and rule out consideration of differences in outcomes across regions. Here, it is useful to distinguish between (1) what is identified, and (2) the sources of identifying variation. The inclusion of region-specific intercepts will often leave the economic interpretation of the slope parameters unchanged, but will use

[10] This perspective acknowledges that, in the growth setting, the true data generating process is unlikely to be among the models considered, even when the number of candidate models is very large. This raises the question of whether AIC-based methods for model averaging are worth exploring; see (Burnham and Anderson, 2002, 2004) and (Claeskens and Hjort, 2008).

within-region variation to identify them.[11] If a researcher remains interested in explaining differences in outcomes across regions, the slope parameters can be combined with the explanatory variables to shed light on this question. Easterly and Levine (1997, Tables VII.A and VII.B) is an example of this approach.

To examine the specific implications for CJ's findings, we consider the following experiment. We investigate what CJ would have found, if they required every growth model to include initial GDP per capita and a small number of region fixed effects. We find that this would have reduced the number of growth determinants found to be robust, and would have drastically reduced the instability of posterior inclusion probabilities across different vintages of the PWT. Moreover, these results continue to apply when we go further than CJ, and consider vintages of the PWT that were not available to them at the time of their study. Hence, we show that modifying the set of candidate models—in particular, reducing the degree of agnosticism, by requiring some variables to appear in each model—should help to limit the number of false positives that might otherwise emerge.

We will present this result in several different ways. As in CJ, our focus is not on the interpretation of the posterior inclusion probabilities of specific regressors, and we are not trying to derive new findings on the determinants of growth. Rather, we examine whether conditioning on initial GDP per capita and regional fixed effects helps to limit instability in the posterior inclusion probabilities. Our first yardstick will be the number of variables which are 'robust', those where the posterior inclusion probability exceeds the prior inclusion probability; this criterion is used in Sala-i-Martin et al. (2004). How does the set of robust variables change across PWT vintages? More precisely, how many variables are robust for at least one PWT vintage, but not all?

These numbers are reported in Table 4.[12] We report two sets of results, one based on the same three vintages used in CJ, and the other augmenting their dataset with the later PWT versions 6.3, 7.0, 7.1, and 8.0. We use the BRIC prior of Fernández et al. (2001b) for the coefficients within each model, and a Bernoulli prior over models where the expected model size is seven variables, as in Sala-i-Martin et al. (2004). The 'CJ' column shows the large number of variables that emerge as robust in one PWT vintage but not in another; this is one reflection of the instability that concerns CJ. The next column restricts attention to conditional convergence regressions (those conditioning on initial GDP per capita, Y_0). This reduces the instability, but less than might have been expected.

Next, we add a small set of region fixed effects: to maintain comparability with previous work, we use the regional dummies that were included in the Sala-i-Martin et al. (2004) analysis. These are dummies for sub-Saharan Africa, East Asia, Europe, and Latin America. The results are stark: the number of variables which are robust at least once, but not always, is reduced by nearly 80%. Allowing for the identity of colonial powers (the final column) results in little further improvement, and for vintages 6.0–8.0, actually increases instability.

Taken together, these results point clearly in one direction. Growth studies are more likely to deliver reliable findings if the candidate models are required to include GDP per capita and region fixed effects. We now provide an alternative summary of the results, which gives more insight into the extent and nature of the improvements in stability. By studying the distribution of the changes in posterior inclusion probabilities, we find evidence that the inclusion of initial GDP capita matters more than Table 4 implies. But the most important finding is that, by including region fixed effects, stability across vintages is greatly improved, because fewer growth determinants are ever assigned a high posterior inclusion probability.

One reason that a more detailed presentation is useful is that posterior inclusion probabilities could display relatively small changes across vintages, and thereby give rise to the apparent lack of robustness highlighted in Table 4. CJ are careful to address this, and show that their results are not simply driven by modest changes either side of thresholds. In our investigation, we now consider a summary measure *MAXDIFF*(z). This is defined, for each candidate explanatory variable z, as the difference between the maximum and minimum posterior inclusion probability across vintages of the PWT

$$MAXDIFF(z) \equiv \underset{i}{Max}\ PIP_i(z) - \underset{j}{Min}\ PIP_j(z)$$

where i and j index vintages of the PWT. By construction, *MAXDIFF*(z) is bounded between zero and one. A high value for a given variable suggests that the posterior inclusion probability of that variable is especially sensitive to the data employed; a value close to zero suggests the posterior inclusion probabilities are relatively stable. Hence, a distribution of *MAXDIFF*(z) that has more mass close to zero indicates a relatively stable set of findings.

The top-left panel of Fig. 1 plots the empirical cumulative distribution function (CDF) of *MAXDIFF*(z) for the results reported by CJ in their Table 2. The largest observed values of *MAXDIFF*(z) are surprisingly close to one, which might be explained by collinearity between, for example, different measures of population density or dummy variables. It is not clear how general this explanation can be, however. The 75th percentile of *MAXDIFF*(z) is close to 0.25, and so a major change in posterior inclusion probabilities across vintages—a change larger than 0.25—is observed for more than a quarter of the candidate growth determinants.

The top-right panel repeats this analysis, but this time requiring initial GDP per capita Y_0 to be included in all the models. Now the maximum of *MAXDIFF*(z) is below 0.75 and, more importantly, the 75th percentile is only 0.09: so for three-quarters of the variables, the change in the posterior inclusion probability across vintages is now less than 0.10. The lower

[11] In much the same way, panel data estimation of the Solow model with country fixed effects (for example, Islam, 1995) uses within-country variation to identify the same structural parameters as the cross-section study of Mankiw et al. (1992).

[12] These and the other BMA results in the paper were obtained using the BMS software due to Feldkircher and Zeugner.

Table 4
Model Stability Comparisons.[a]

Datasets	CJ	Y_0	Y_0& Continents	Y_0& Continents & Colonial powers
PWT 6.0–6.2	19	15	4	3
PWT 6.0–8.0	24	22	6	7

[a] This table reports the number of variables that are 'robust' (see text) for at least one vintage of the PWT but not all. Column CJ refers to the approach of Ciccone and Jarociński (2010). The next columns require all models to include initial GDP per capita, Y_0; initial GDP per capita and four region dummies; initial GDP per capita, four region dummies, and dummies for colonial powers. The first row uses the three vintages of the PWT considered by CJ; the second adds four subsequent vintages, up to the PWT 8.0 release.

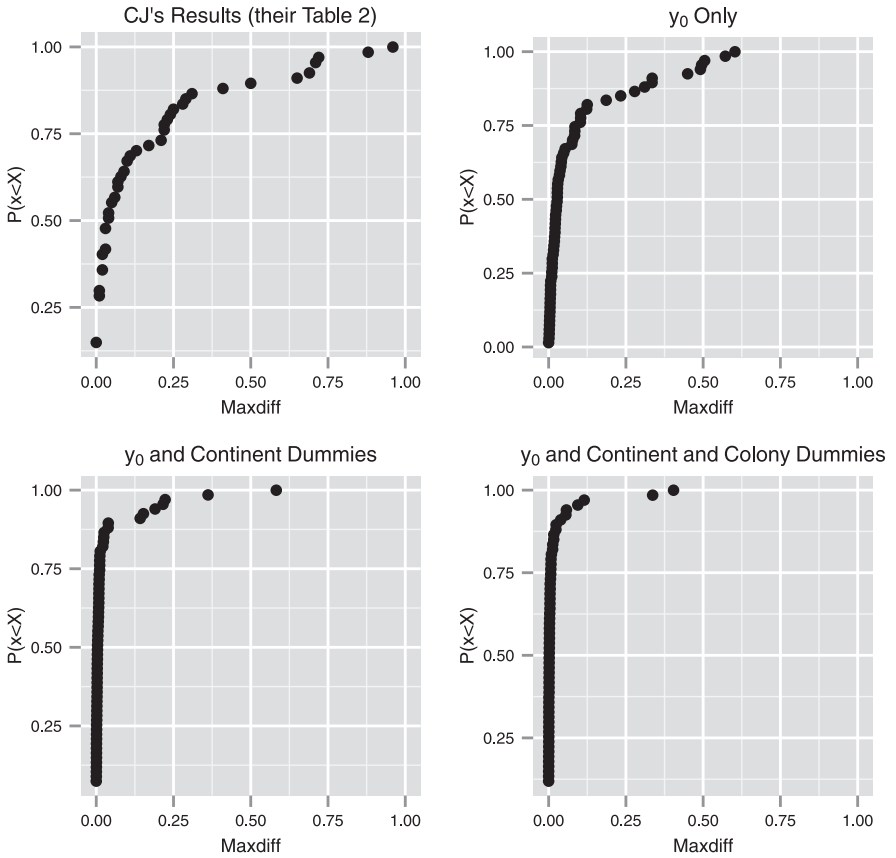


Fig. 1. Maximum difference in posterior inclusion probabilities across PWT 6.0–6.2. Figures report the empirical CDF of the maximum difference in PIPs across PWT vintages.

panels add region fixed effects (bottom-left) and also dummies for the former colonial power (bottom-right). These panels show a pronounced reduction in the number of large changes, with the CDF moving much closer towards the y-axis.

Given that later vintages of the PWT have been released since CJ wrote their paper, we ask whether extending their analysis to these additional vintages gives similar results. The top-left panel of Fig. 2 displays the empirical CDFs of *MAXDIFF* (z) based on differences across four newer vintages of the PWT, 6.3, 7.0, 7.1 and 8.0. These results are quite close to those in the top-left panel of Fig. 1. But again, the instability is reduced by the inclusion of initial GDP per capita and region fixed effects in all models. The main departure from the previous results is that including dummies for the former colonial power (the bottom-right panel) is a more noticeable improvement.

Fig. 3 considers the same four cases as Fig. 2, but now *MAXDIFF*(z) is calculated across seven vintages of the PWT. By construction, considering a larger number of vintages cannot reduce *MAXDIFF*(z). Put differently, on this metric, the instability that CJ identify can only get worse as the PWT data are repeatedly updated and revised, unless a researcher can confidently identify some vintages as more reliable than others. But Fig. 3 shows the same overall pattern: instability is sharply reduced by forcing the inclusion of initial GDP capita, region fixed effects, and dummy variables for former colonial powers.

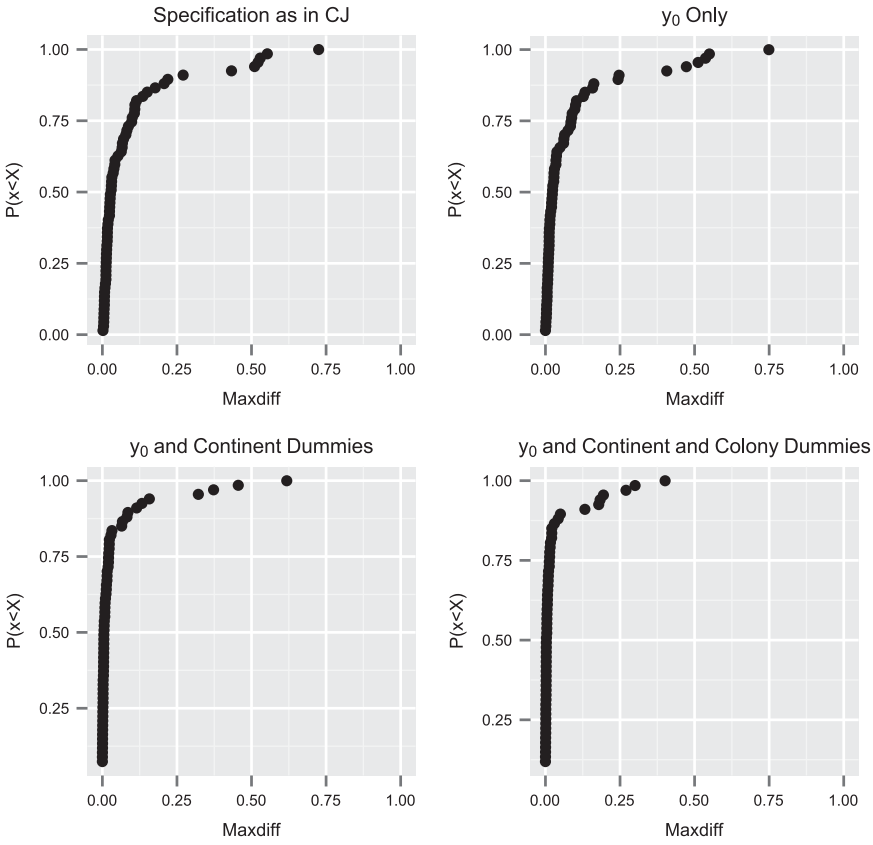


Fig. 2. Maximum difference in posterior inclusion probabilities across PWT 6.3–8.0.

We now briefly investigate a related issue: as Feldkircher and Zeugner (2012) emphasize, the posterior inclusion probabilities are unstable across vintages in CJ partly because the set of countries varies. The countries which move in and out of the sample are disproportionately poor and/or in sub-Saharan Africa. With this in mind, we repeat our experiment using a fixed subsample of 75 countries, initially studying vintages 6.0, 6.1, and 6.2. We use a new style of figure, which also plots the minimum and maximum posterior inclusion probability for each variable, linking the extremes with a horizontal bar for ease of interpretation. This is shown in Fig. 4, with a vertical line at a posterior inclusion probability of $x = 7/67$, representing the Sala-i-Martin et al. (2004) criterion for deeming a result to be of interest. The long horizontal bars in the top-left panel show the considerable instability in the posterior inclusion probabilities highlighted by CJ. The seven variables identified as robust by CJ are those for which both the maximum and minimum are to the right of the vertical line.

The other panels show the difference made by a fixed sample (top-right), the inclusion of initial GDP per capita (bottom-left) and the inclusion of initial GDP per capita and region fixed effects (bottom-right). In the latter case, there is an especially clear partition of the variables into those which potentially matter, and those which appear in relatively few well-performing models: most of the horizontal bars are much shorter than before. This suggests that, as anticipated, the inclusion of region fixed effects helps to guard against potential false positives.

A remaining question is whether any variables emerge as consistently important. Table 5 shows the minimum, median and maximum posterior inclusion probability for the four variables whose posterior inclusion probability exceeds the 7/67 threshold across three vintages of the PWT, when initial GDP and four region fixed effects are included. In several cases, the variation emphasized by CJ remains visible. But the evidence that primary schooling in 1960 matters is strong: its posterior inclusion probability is close to unity across three vintages of the PWT. Put differently, those models which exclude initial primary schooling are all assigned very low posterior model probabilities. Given that the benefits of education for growth have often been contested, this is a striking result.

The remaining problem is that these findings do not extend to more recent vintages of the PWT. We repeat the above analysis for the four subsequent vintages of the PWT available to us, with results reported in Table 6. The fixed sample of countries differs, but otherwise, only the vintages of the PWT vary between the two tables. The variables which emerge as consistently 'robust' are now somewhat different, and the variables previously identified as robust (shown in the last four rows) now have rather variable inclusion probabilities. In terms of substantive conclusions about growth, the alternative sets of variables listed in Tables 5 and 6 point to factors that are widely associated with East Asia's rapid growth — even though we are controlling for an East Asian dummy in these models. Put differently, past emphasis on growth-promoting

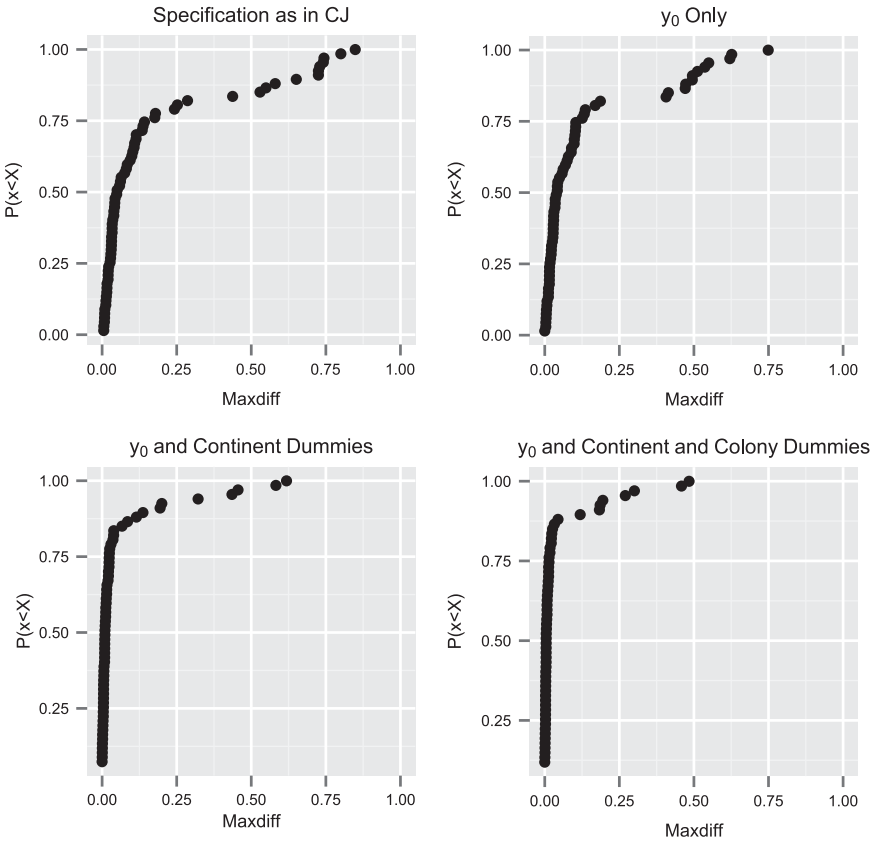


Fig. 3. Maximum difference in posterior inclusion probabilities across PWT 6.0–8.0.

aspects of East Asia's initial conditions and subsequent policies, such as high primary schooling, low specialization in primary exports/mining, and relatively open trade policies, may have been on the right track. But as Table 6 indicates, it is difficult to infer which of these forces matter most, as CJ would have anticipated.

It is clear that when initial GDP per capita and region fixed effects are always included in the models, the list of other growth determinants consistently identified by BMA is short, especially compared to the many candidates that have been considered. Our findings in this regard are consistent with CJ's overall conclusion, quoted earlier in this paper, that the data uncertainties may be too great to identify more than a small number of growth determinants. Researchers who begin as agnostics may be destined to remain as agnostics.

In their own extension of the CJ results, Feldkircher and Zeugner (2012) similarly conclude in favour of 'robust ambiguity'. They find that, when using a hierarchical prior which allows more flexibility in the extent of shrinkage, the posterior mass is distributed more evenly across candidate models. This reduces the variation of posterior inclusion probabilities across vintages of the PWT data, but also reduces the variation across variables for a given vintage. That finding raises the following question: if we combine their priors with the ideas of this paper, and restrict the model space, can we start to discriminate more readily between variables?

To investigate this, we follow them in using a uniform model prior, but a flexible 'hyper-*g*' prior for the coefficients in each model. This allows the extent of shrinkage to be influenced by the data, and should make the analysis less dependent on prior assumptions, which are inevitably arbitrary to some degree. Hierarchical priors of this general form are among those considered by Feldkircher and Zeugner (2009), Ley and Steel (2012), and Liang et al. (2008). Ley and Steel (2012) provide further references, and note that the effect is to extend the conventional *g*-prior to be a mixture of normals, with more flexible tails than a normal distribution. For comparability with Feldkircher and Zeugner (2012), we set the expected extent of shrinkage (the mean of the prior over the '*g*' parameter) equal to the (fixed) extent of shrinkage found in the benchmark BRIC prior of Fernández et al. (2001b).

The results for version 8.0 of the PWT are shown in Table 7. Here, we are less interested in variability across vintages of the data. Instead, we want to know how the posterior inclusion probabilities vary, under flexible priors, or flexible priors with some variables always included. With this in mind, the table reports some of the percentiles of the PIP distribution, and the interquartile range (IQR). The table shows that forcing the models to include initial GDP per capita slightly raises the interquartile range; then, additionally forcing the models to include region fixed effects lowers the interquartile range substantially. But the latter effect arises because many of the posterior inclusion probabilities are moved closer to zero. The

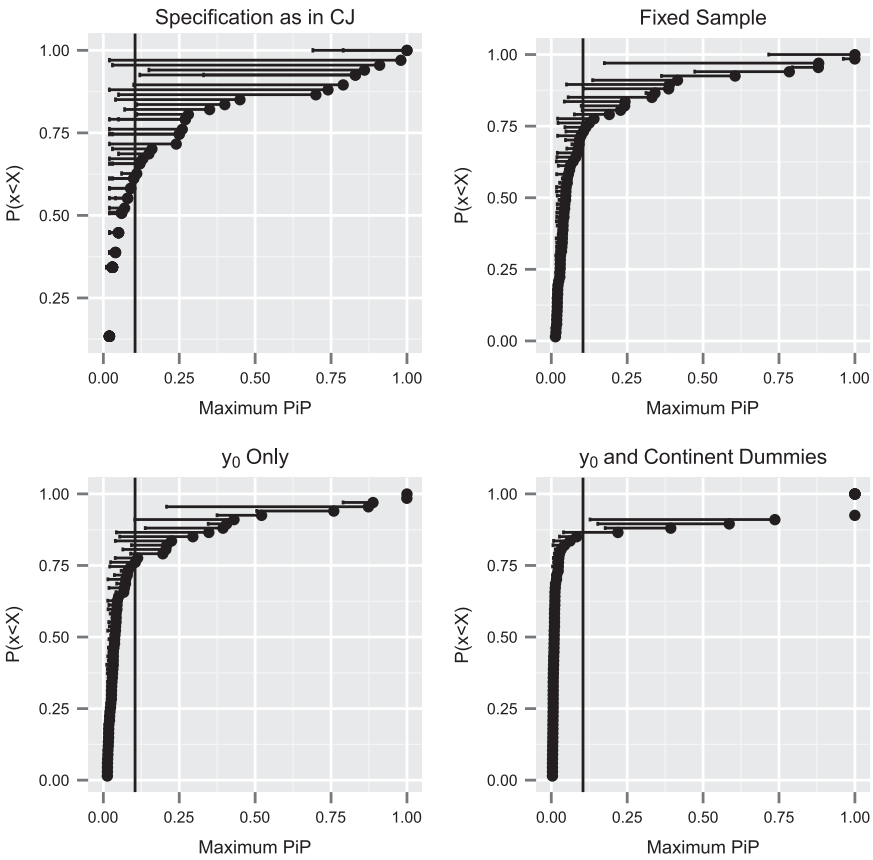


Fig. 4. Posterior inclusion probabilities across PWT Vintages 6.0–6.2. The figures show the empirical CDF of the maximum PIP across PWT vintages. In addition, horizontal bars extend between the minimum and maximum PIP for each variable. The vertical line, $x = 7/67$, is the Sala-i-Martin et al. (2004) threshold for a variable to be considered robust.

Table 5
PIPs for robust variables based on fixed sample PWT 6.0–6.2.[a]

Variable	Minimum PIP	Median PIP	Maximum PIP
Primary Schooling in 1960	0.99	1.00	1.00
Fertility 1960s	0.13	0.61	0.74
Primary Exports 1970	0.15	0.43	0.59
Fraction Confucius	0.18	0.27	0.39

Table 6
PIPs for robust variables based on fixed sample PWT 6.3—8.0.[a]

Variable	Minimum PIP	Median PIP	Maximum PIP
Life expectancy in 1960s	0.34	0.63	0.78
Mining as a percentage of GDP	0.17	0.94	0.99
Openness 1965–1974	0.17	0.31	0.43
Primary schooling in 1960	0.08	0.37	0.56
Fertility 1960s	0.07	0.14	0.24
Primary Exports 1970	0.07	0.17	0.21
Fraction Confucius	0.03	0.19	0.25

number of candidate variables with an inclusion probability greater than 7/67 (see the final row of the table) is now rather small. Similar results apply for earlier vintages of the PWT data (results not reported). These results show that our main points—the need to include initial GDP per capita, and region dummies—have implications for the results even when a

Table 7
Percentiles for PIP distribution—PWT 8.0.[a]

Percentile	Hyper-*g*	Hyper-*g* & Y0	Hyper-*g* & Y0 & Continents
0.25	0.06	0.05	0.02
0.50	0.08	0.06	0.03
0.75	0.15	0.14	0.05
0.90	0.26	0.24	0.15
IQR	0.085	0.090	0.032
$No.PIP \geq 7/67$	20	18	7

[a] This table shows the percentiles of the PIP distribution for PWT 8.0, when a hyper-*g* prior is used, and then when initial GDP per capita is always included, and then when initial GDP per capita and region dummies are always included. In each case, the PIP distribution is for a fixed set of variables, excluding GDP per capita and the region dummies.

hyperprior is used; and that the method does isolate a limited number of variables as more important than others. The results also continue to suggest, in line with CJ, that most of the variables do not have much support from the data.

This does not mean that nothing matters, however. Interestingly, in these results based on a hyper-*g* prior, the one variable which repeatedly emerges with a high posterior inclusion probability is the primary school enrollment rate in 1960. In one sense, this is not a new result; it is one of the variables emphasized most by Sala-i-Martin et al. (2004), given its high posterior inclusion probability in their study. Before them, Levine and Renelt (1992) had identified an alternative measure of mass education—the secondary school enrollment rate—as an unusually robust determinant of growth. Now, with additional vintages of data, and improved methods for addressing model uncertainty, it seems clear that both sets of authors were right to draw attention to schooling. It is especially important given the long-standing debate over whether donors and developing-country governments should prioritize growth or meeting basic needs; broad-based schooling can address both.[13] That seems well worth knowing, and in itself, worth the energy that has been expended upon these methods.

5. Discussion

We now discuss some broader issues. For its detractors, the empirical growth literature is often said to ask questions of the data that the data are ill-equipped to answer. The critics have found support, at least indirectly, in the work of Jerven (2013). He points to the many problems with national accounts statistics in sub-Saharan Africa, raising doubts about what can be learnt from the data. It is interesting to ask whether the issues identified by CJ are primarily the weaknesses of (one version of) BMA, or weaknesses that arise from mismeasurement.

Perhaps inevitably given the complexity of the exercise, at least some versions of the PWT have contained errors. Jerven (2013, pp. 70–71) discusses a major problem with the PWT 6.1 data on GDP for Tanzania, while Dowrick (2005) uncovered errors in PWT 6.1 data on labour force participation rates for around 15 countries in sub-Saharan Africa. More fundamentally, the compilers of the PWT have had to face difficult choices about how to use the price data from multiple benchmark years, and how to construct PPP estimates for countries where ICP benchmark data are not available. In practice, this means that the data revisions across PWT vintages can be significant, as Johnson et al. (2013) analyze in detail. But they also find that the conclusions of some prominent studies are quite robust, especially when studies are based on growth rates calculated over long spans of time, rather than making intensive use of the high-frequency variation. They conclude 'whatever else is right or wrong with the growth literature, the bulk of it is not afflicted by the problem of sensitivity to changes in PWT GDP data' (Johnson et al., 2013, p. 266).

Their findings suggest that the instability of BMA cannot be attributed solely to weaknesses in the data. It remains possible that BMA, in the forms implemented here, is too sensitive to small changes in the sample, perhaps because of the effects of outliers. In the case of a BIC-based approach, this would not be a surprise: in the normal linear model, the *BIC* can be written as a function of the R^2, and the R^2 is a highly non-robust statistic. More generally, most Bayesian approaches to model averaging rely on the sum of squared errors as the measure of fit, and this is similarly non-robust. The simulations in Machado (1993) indicated that fat-tailed error distributions induce a tendency for the *BIC* to select models which are too small. Machado introduced an approach for making the *BIC* outlier-robust; see also (Claeskens and Hjort, 2008, pp. 81–82). In principle, the weights used in model averaging could be based on statistics that are more robust than the *BIC*, and this idea seems worth pursuing in future work. So does the simpler approach of drawing conclusions based on repeatedly taking subsamples from the data. This method is adopted in Ley and Steel (2009, 2012) and should arguably become standard in future studies that use model averaging.

It is worth noting that outlier-robustness is a consideration even with a fixed sample, because the influence exerted by outliers will change with each model considered. If we think of outliers as observations that are generated from a distinct data-generating process, perhaps one of secondary interest, the extent to which these observations are well approximated

[13] It could be objected that this is only a statistical association, not a causal effect, and that Bils and Klenow (2000) provide an alternative explanation for the association. But views will differ on its plausibility, and it seems hard to argue that someone new to this literature should not update their priors.

will vary across models. This raises the possibility that certain models have high posterior model probabilities mainly because a particular combination of variables captures data variation that, in a more robust analysis, would have been identified as atypical. Ideally, variable selection or model averaging require the simultaneous consideration of outliers, especially in samples of the size considered here; see Temple (2000) for further discussion and references.

This can be extended to a broader criticism of model averaging as usually implemented by economists. Statisticians have often made the point that model criticism or model checking is a fundamental component of empirical research. In contrast, most of the applications of BMA by economists are a little mechanical in this regard, with limited use of plots of the data, or careful study of the properties of individual models. Our aim has been to study the sensitivity of BMA rather than derive a set of substantive conclusions about growth; a careful investigation of the latter would need to take a more flexible, iterative approach to aspects of the data and the specification.[14]

This relates to another broad point, namely the underlying aims of using BMA. One reason for using Bayesian approaches is that conventional significance tests are rarely a sound basis on which to form policy or take decisions, as Brock and Durlauf (2001) and Brock et al. (2003) emphasized.[15] In this paper, we have followed much of the growth literature in emphasizing posterior inclusion probabilities. One justification is that a theorist might be interested in knowing which variables appear to influence growth, while taking a conservative approach that is wary of 'false positives'. For policy-makers, however, the posterior distributions of the parameters are especially relevant, arguably more so than the posterior inclusion probabilities. It would be possible for a posterior distribution to have substantial mass at zero, but also to assign sufficient mass to large negative or positive effects (or both) that the variable becomes highly relevant to a policy decision.

This suggests that an especially interesting sequel to Ciccone and Jarociński would be to reframe the problem in terms of a policy decision, and examine its sensitivity to the use of alternative vintages of data. We leave this as an idea for further work, but note that once policy decisions become the central focus, it is less clear that region fixed effects should be included in every model considered, and one might also want to assign at least a little prior probability to the existence of long-run growth effects.

6. Conclusions

At least in economics, it seems likely that many published findings are based at least partly on data mining and 'directed' model searches. Even those researchers who adopt a more agnostic, open-minded approach rarely take model selection into account when carrying out inference. These problems suggest that model selection, estimation, and inference should be considered jointly. Over the last twenty years, Bayesian Model Averaging has emerged as an attractive approach, and one which is increasingly easy to implement. In this paper, we have examined one of the most important critiques of BMA, that advanced by Ciccone and Jarociński (2010).

We revisit their theoretical arguments, and note some additional considerations that qualify the arguments a little. We also reconsider their empirical analysis. In particular, we show that instability in posterior inclusion probabilities is sharply reduced by restricting the set of candidate models, so that initial GDP per capita and a small set of region fixed effects are included in each model. There are theoretical arguments for requiring these particular variables to be included. Moreover, we find that their inclusion enhances robustness even when we extend the analysis to more recent vintages of the Penn World Table, not studied by CJ. For many variables, the posterior inclusion probabilities do not vary widely across vintages of the PWT, calling into question the generality of some of the points made by CJ. But in closer agreement with CJ, we also find that few variables are consistently identified as important growth determinants. Our restriction of the model space works mainly because high posterior inclusion probabilities become less common.

As this suggests, many of the points made by CJ have some force. Even using the approach we have recommended in this paper, there remains some instability across data vintages, suggesting that they were correct to raise concerns. It is likely that future research will draw on newer methods, especially more flexible priors, such as those advocated by Feldkircher and Zeugner (2009) and Ley and Steel (2012). In the meantime, attention to the set of candidate models is likely to bring significant benefits; one can have too much agnosticism. Our results suggest that cross-section growth regressions should routinely include initial GDP per capita and region fixed effects, whether or not model averaging is used.

[14] On Bayesian methods for model checking and sensitivity analysis, see chapters 6 and 7 of Gelman et al. (2014). Gelman and Shalizi (2013) discuss the place of these methods within a wider philosophy of statistics.

[15] A principal theme of McGrayne (2011) is that Bayesian principles have often been successful in practical applications for which classical methods are either ill-equipped or actively misleading. For accounts of decision theory linked to Bayesian principles, see chapter 9 of Gelman et al. (2014), or (Leonard and Hsu, 1999). Kass (2011) gives an account of the pragmatism of modern statistical practice.

Appendix A

We briefly discuss the penalty for loss of parsimony embedded in the BIC approach. Using the definition of the adjusted R^2 as

$$\overline{R}_q^2 \equiv 1 - \left(\frac{N-1}{N-q-1}\right)\left(1-R_q^2\right)$$

we can rewrite the posterior inclusion probability (for the normal linear model, with one model per size class) in terms of each dominant model's adjusted R^2

$$PIP(z) \propto \sum_{q=1}^{Q} (1-\overline{R}_q^2)^{-N/2}\omega_q P_q \Lambda_q(z)$$

where the weights ω_q are

$$\omega_q \equiv \left(1-\frac{1+q}{N}\right)^{-N/2}\left(1-\frac{2}{N}\right)^{N/2} N^{-(q-1)/2}$$

For $N=88$, the sample size in Sala-i-Martin et al. (2004), the weights decline rapidly, in the ratio of approximately 5:1. In other words, in computing the posterior inclusion probability, the BIC approach will not give much weight to larger models unless they increase the adjusted R^2 substantially. This reflects a well-known property of the BIC penalty for model size: it is much stricter than the penalty embedded in the adjusted R^2. Kuha (2004, p. 216) gives an especially clear account of the reasons for favouring parsimony.

Appendix B. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.euroecorev.2015.05.010.

References

Amini, S., Parmeter, C.F., 2012. Comparison of model averaging techniques: assessing growth determinants. J. Appl. Econom. 27, 870–876.

Banerjee, A.V., 2009. Big answers for big questions: the presumption of growth policy. In: J. Cohen, W. Easterly (Eds.), What Works in Development? Thinking Big and Thinking Small, Brookings Institution Press, Washington DC.

Barro, R.J., 1991. Economic growth in a cross section of countries. Q. J. Econ. 106 (2), 407–443.

Barro, R.J., Sala-i-Martin, X., 1992. Convergence. J. Polit. Econ. 100 (2), 223–251.

Bartelsman, E.J., Wolf, Z., 2014. Forecasting aggregate productivity using information from firm-level data. Rev. Econ. Stat. 96, 745–755.

Baumol, W.J., 1986. Productivity growth, convergence, and welfare: what the long-run data show. Am. Econ. Rev. 76 (5), 1072–1085.

Begun, J., Eicher, T., 2008. In search of a sulphur dioxide environmental Kuznets curve: a Bayesian Model Averaging approach. Environ. Dev. Econ. 13 (6), 1–28.

Bils, M., Klenow, P.J., 2000. Does schooling cause growth? Am. Econ. Rev. 90 (5), 1160–1183.

Breiman, L., 1992. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. J. Am. Stat. Assoc. 87 (419), 738–754.

Brock, W.A., Durlauf, S.N., 2001. Growth empirics and reality. World Bank Econ. Rev. 15 (2), 229–272.

Brock, W.A., Durlauf, S.N., West, K.D., 2003. Policy evaluation in uncertain economic environments. Brook. Pap. Econ. Act. 34 (1), 235–322.

Brodeur, A., Lé, M., Sangnier, M., Zylberberg, Y., 2015. Star wars: the empirics strike back, Manuscript, Paris School of Economics, March, Am. Econ. J.: Appl. Econ. (forthcoming).

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference, second edition Springer, New York.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociol. Methods Res. 33, 261–304.

Ciccone, A., Jarociński, M., 2010. Determinants of economic growth: will data tell?. Am. Econ. J.: Macroecon. 2 (4), 222–246.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge.

Cohen-Cole, E.B., Durlauf, S.N., Rondina, G., 2012. Nonlinearities in growth: from evidence to policy. J. Macroecon. 34, 42–58.

Crainiceanu, C.M., Dominici, F., Parmigiani, G., 2008. Adjustment uncertainty in effect estimation. Biometrika 95 (3), 635–651.

Crespo Cuaresma, J., Doppelhofer, G., Feldkircher, M., 2014. The determinants of economic growth in European regions. Reg. Stud. 48, 44–67.

Danquah, M., Moral-Benito, E., Ouattara, B., 2014. TFP growth and its determinants: a model averaging approach. Empir. Econ. 47, 227–251.

Deckers, T., Hanck, C., 2014. Variable selection in cross-section regressions: comparisons and Extensions. Oxf. Bull. Econ. Stat. 76, 841–873.

Doppelhofer, G., Weeks, M., 2009. Jointness of growth determinants. J. Appl. Econom. 24 (2), 209–244.

Doppelhofer, G., Weeks, M., 2011. Robust Growth Determinants, CESifo Working Paper series No. 3354.

Dowrick, S., 2005. Errors in the Penn World Table demographic data. Econ. Lett. 87, 243–248.

Durlauf, S.N., Fu, C., Navarro, S., 2012. Assumptions matter: model uncertainty and the deterrent effect of capital punishment. Am. Econ. Rev. Pap. Proc. 102 (3), 487–492.

Easterly, W., 2009. Comment on Banerjee. In: J. Cohen, W. Easterly (Eds.), 2009. What Works in Development? Thinking Big and Thinking Small, Brookings Institution Press, Washington DC.

Easterly, W., Levine, R., 1997. Africa's Growth Tragedy: Policies and Ethnic Divisions. Q. J. Econ. 112(4), 1203–1250.

Eicher, T.S., Henn, C., Papageorgiou, C., 2012. Trade creation and diversion revisited: accounting for model uncertainty and natural trading partner effects. J. Appl. Econ. 27, 296–321.

Eicher, T., Papageorgiou, C., Raftery, A.E., 2010. Determining growth determinants: default priors and predictive performance in Bayesian Model Averaging. J. Appl. Econ. 26, 30–55.

Eicher, T.S., Papageorgiou, C., Roehn, O., 2007. Unraveling the fortunes of the fortunate: an Iterative Bayesian Model Averaging (IBMA) approach. J. Macroecon. 29, 494–514.

Fanelli, D., 2010. 'Positive' results increase down the hierarchy of the sciences. PLoS ONE 5 (4), 1–10.

Feldkircher, M., 2014. The determinants of vulnerability to the global financial crisis 2008 to 2009: credit growth and other sources of risk. J. Int. Money Finance 43, 19–49.
Feldkircher, M., Horvath, R., Rusnak, M., 2014. Exchange market pressures during the financial crisis: a Bayesian Model Averaging evidence. J. Int. Money Finance 40, 21–41.
Feldkircher, M., Zeugner, S., 2009. Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging. IMF Working Papers, 09/202.
Feldkircher, M., Zeugner, S., 2012. The impact of data revisions on the robustness of growth determinants - a note on 'Determinants of economic growth: will data tell?'. J. Appl. Econ. 27, 686–694.
Fernández, C., Ley, E., Steel, M.F.J., 2001a. Model uncertainty in cross-country growth regressions. J. Appl. Econ. 16 (5), 563–576.
Fernández, C., Ley, E., Steel, M.F.J., 2001b. Benchmark priors for Bayesian Model Averaging. J. Econ. 100 (2), 381–427.
Garratt, A., Lee, K., 2010. Investing under model uncertainty: decision based evaluation of exchange rate forecasts in the US, UK and Japan. J. Int. Money Finance 29 (3), 403–422.
Gassebner, M., Lamla, M.J., Vreeland, J.R., 2013. Extreme Bounds of Democracy. J. Confl. Resolut. 57, 171–197.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. Bayesian Data Analysis, CRC Press third edition Taylor and Francis, Boca Raton, FL.
Gelman, A., Rubin, D.B., 1995. Avoiding model selection in Bayesian social research. Sociol. Methodol., 165–173.
Gelman, A., Shalizi, C.R., 2013. Philosophy and the practice of Bayesian statistics. Br. J. Math. Stat. Psychol. 66, 8–38.
George, E.I., Foster, D.P., 2000. Calibration and empirical Bayes variable selection. Biometrika 87 (4), 731–747.
González-Aguado, C., Moral-Benito, E., 2013. Determinants of corporate default: a BMA approach. Appl. Econ. Lett. 20 (6), 511–514.
Gottardo, R., Raftery, A.E., 2009. Bayesian robust variable and transformation selection: a unified approach. Can. J. Stat. 37, 1–20.
Hansen, B.E., 2007. Least squares model averaging. Econometrica 75 (4), 1175–1189.
Hansen, B.E., 2014. Model averaging, asymptotic risk, and regressor groups. Quant. Econ. 5, 495–530.
Hendry, D.F., Krolzig, H.-M., 2004. We Ran One Regression. Oxf. Bull. Econ. Stat. 66 (5), 799–810.
Hlouskova, J., Wagner, M., 2013. The determinants of long-run economic growth: a conceptually and computationally simple approach. Swiss J. Econ. Stat. 149 (4), 445–492.
Huang, Y., 2011. Determinants of Financial Development. Palgrave MacMillan, Basingstoke.
Islam, N., 1995. Growth empirics: a panel data approach. Q. J. Econ. 110 (4), 1127–1170.
Jensen, P.S., Würtz, A.H., 2012. Estimating the effect of a variable in a high-dimensional linear model. Economet. J. 15, 325–357.
Jerven, M., 2013. Poor Numbers. Cornell University Press, Ithaca, NY.
Johnson, S., Larson, W., Papageorgiou, C., Subramanian, A., 2013. Is newer better? Penn World Table revisions and their impact on growth estimates. J. Monet. Econ. 60, 255–274.
Kass, R.E., 2011. Statistical inference: the big picture. Stat. Sci. 26 (1), 1–9.
Koop, G., Leon-Gonzalez, R., Strachan, R., 2012. Bayesian Model Averaging in the instrumental variable regression model. J. Economet. 171 (2), 237–250.
Kormendi, R.C., Meguire, P.G., 1985. Macroeconomic determinants of growth: cross-country evidence. J. Monet. Econ. 16 (2), 141–163.
Kraay, A., Tawara, N., 2013. Can specific policy indicators identify reform priorities? J. Econ. Growth 18, 253–283.
Kuha, J., 2004. AIC and BIC: comparisons of assumptions and performance. Sociol. Methods Res. 33 (2), 188–229.
Leamer, E.E., 1974. False models and post-data model construction. J. Am. Stat. Assoc. 69, 122–131.
Leamer, E.E., 1978. Specification Searches. John Wiley, New York.
Leamer, E.E., 1983. Let's take the con out of econometrics. Am. Econ. Rev. 73 (1), 31–43.
Leamer, E.E., Leonard, H., 1983. Reporting the fragility of regression estimates. Rev. Econ. Stat. 65 (2), 306–317.
Leonard, T., Hsu, J.S.J., 1999. Bayesian Methods. Cambridge University Press, Cambridge.
Levine, R., Renelt, D., 1992. A sensitivity analysis of cross-country growth regressions. Am. Econ. Rev. 82 (4), 942–963.
Ley, E., Steel, M.F.J., 2007. Jointness in Bayesian variable selection with applications to growth regression. J. Macroecon. 29 (3), 476–493.
Ley, E., Steel, M.F.J., 2009. On the effect of prior assumptions in Bayesian Model Averaging with applications to growth regression. J. Appl. Econ. 24 (1), 651–674.
Ley, E., Steel, M.F.J., 2012. Mixtures of *g*-priors for Bayesian Model Averaging with economic applications. J. Econ. 171, 251–266.
Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O. (2008). Mixtures of *g*-priors for Bayesian variable selection. J. Am. Stat. Assoc. 103 (481), 410–423.
Machado, J.A.F., 1993. Robust model selection and M-estimation. Econ. Theory 9, 478–493.
Magnus, J.R., De Luca, G., 2014. Weighted-average least squares (WALS): a survey. J. Econ. Surv., forthcoming.
Magnus, J.R., Powell, O., Prüfer, P., 2010. A comparison of two model averaging techniques with an application to growth empirics. J. Econ. 154 (2), 139–153.
Magnus, J.R., Wang, W., 2014. Concept-based Bayesian Model Averaging and growth empirics. Oxf. Bull. Econ. Stat. 76, 874–897.
Malik, A., Temple, J.R.W., 2009. The geography of output volatility. J. Dev. Econ. 90 (2), 163–178.
Mankiw, N.G., Romer, D., Weil, D., 1992. A contribution to the empirics of economic growth. Q. J. Econ. 107 (2), 407–437.
Marris, R., 1982. How much of the slow-down was catch-up?. In: Matthews, R.C.O. (Ed.), Slower Growth in the Western World, Heinemann, London.
Masanjala, W.H., Papageorgiou, C., 2008. Rough and lonely road to prosperity: a reexamination of the sources of growth in Africa using Bayesian Model Averaging. J. Appl. Econom. 23, 671–682.
McGrayne, S.B., 2011. The Theory That Would Not Die. Yale University Press, New Haven.
Mitchell, J., Pain, N., Riley, R., 2011. The drivers of international migration to the UK: a panel-based Bayesian Model Averaging approach. Econ. J. 121, 1398–1444.
Moral-Benito, E., 2012. Determinants of economic growth: a Bayesian panel data approach. Rev. Econ. Stat. 94 (2), 566–579.
Raftery, A.E., 1995. Bayesian model selection in social research (with discussion). Sociol. Methodol. 25, 111–196.
Raftery, A.E., Madigan, D., Hoeting, J.E., 1997. Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. 92, 179–191.
Sala-i-Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of long-term growth: a Bayesian Averaging of Classical Estimates (BACE) approach. Am. Econ. Rev. 94 (4), 813–835.
Salimans, T., 2012. Variable selection and functional form uncertainty in cross-country growth regressions. J. Econom. 171, 267–280.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Sirimaneetham, V., Temple, J., 2009. Macroeconomic stability and the distribution of growth rates. World Bank Econ. Rev. 23 (3), 443–479.
Strachan, R.W., van Dijk, H.K., 2013. Evidence on features of a DSGE Business Cycle Model from Bayesian Model Averaging. Int. Econ. Rev. 54 (1), 385–402.
Surowiecki, J., 2004. The Wisdom of Crowds. Doubleday, New York.
Temple, J., 1998. Equipment investment and the Solow model. Oxf. Econ. Pap. 50 (January (1)), 39–62.
Temple, J., 2000. Growth regressions and what the textbooks don't tell you. Bull. Econ. Res. 52 (3), 181–205.
Temple, J., 2003. The long-run implications of growth theories. J. Econ. Surv. 17 (3), 497–510.
Wright, J.H., 2008. Bayesian Model Averaging and exchange rate forecasts. J. Econom. 146 (2), 329–341.