The stringent, the liberal, and the robust:

Can specification search algorithms help us solve the cross-country growth puzzle?

Willem van der Deijl
Erasmus University, Rotterdam

Title thesis: The liberal, the robust and the stringent: Can specification search algorithms help us solve the cross-country growth puzzle?

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Department of Economics

Supervisor: Nalan Baştürk

Name: Willem van der Deijl
Exam number: 362234
E-mail address: willemvanderdeijl@hotmail.com

Abstract

Due to a lack of a unified theoretical framework for empirical research on the determinants of long-term growth, different specification search algorithms have been applied to this field in the past decade. These include the general-to-specific approach (*Gets*), Bayesian Model Averaging (BMA), and Weighted-Average-Least-Squares (WALS). These methods are critically assessed in the context of cross-country growth regressions. Their efficacy to find the correct specification is evaluated by means of a number of Monte Carlo experiments. Robustness with respect to nonlinearities and set of potential regressors is assessed as well. BMA is found to be stringent, but reliable, *Gets* is found to be most powerful, but liberal, and WALS is found to be most robust to size of the potential set of regressors. Evidence is found for a tight relationships between long term growth rates and the following variables: initial income, real exchange rate distortions, years open economy and initial fertility rates.

Acknowledgments

Having studied and written a thesis on the philosophy of econometrics before embarking on this project made writing this thesis exciting, interesting and odd. Having written on econometric inference made actually doing it that much more interesting. I am very thankful for having a had a supervisor that was willing to help me with this project, even though I was sometimes quite unfamiliar with the topic at hand. This included some more extensive aid when it came to programming in statistical software I was not familiar with. I would therefore like to thank Nalan Baştürk, without whom I would not have been able to write this thesis. Furthermore I would like to thank my parents, Leenderd en Christina van der Deijl, for emotional and financial support; Nina Kloeg for helping me in the process, and Viorel Milea for being available as a second reader.

Contents

Contents 4
1. Introduction: 5
2. Motivation and Literature Review 8
2.1 Growth Theory and Open-Endedness 8
2.2. Specification Selection Methods 10
3. The experiment 16
4. Results 24
5. Real growth data 33
7. Summary and conclusion 41
Appendix 42
References 45

1. Introduction:
Growth empirics and cross-country data

Why are some countries rich and other poor? Or, more importantly, how can we devise policies to stimulate development in poorer countries? These questions constitute the main topics of research in the field of growth economics. Unfortunately, finding an answer to these questions is complex. Recent literature in the field of growth economics deals with the issue of open-endedness of growth theories (Brock and Durlauf, 2001): growth theories propose a large number of explanatory variables, but fail to specify precise models, or exclude certain potential explanatory variables of growth. This has adverse consequences in the field, as it leaves the field without a unified formal theory that can guide research. This lack of a unified formal theory causes difficulties for classical approaches to econometrics, because such approaches to econometrics often maintain that data can be used to test theories, but that data cannot be used to form them (explicit in Koopmans, 1947; cf. Hoover and Perez, 2000; Backhouse and Morgan, 2000; Du Plessis, 2009). As a result, the field of growth economics has shifted its focus to less orthodox empirical methods that do not rely so heavily on theory to resolve this issue (see Durlauf et al., 2005 and Ulasan, 2011 for such an argument; notable examples of research using highly data-driven methods are Levine and Renelt, 1992; Sala-i-Martin, 1997; Fernandez et al., 2001b; Sala-i-Martin et al., 2004; Hendry and Krolzig, 2004; Hoover and Perez, 2004; Cuaresma and Doppelhofer, 2007; Eberhardt and Teal, 2011; Salimans, 2012; just to name few). In order to highlight how highly empirical growth economics has become, Durlauf et al. speak of "growth econometrics".

A substantial field of econometrics deals with this issue under the name of "specification searches", "search algorithms", or "(weak) data mining". It is somewhat difficult to delineate these concepts. Few econometrics texts (text books as well as methodological papers) are explicit on the specific relationship they envision between theory and evidence, and what exactly these terms mean. In this paper methods are discussed that use purely statistical arguments to specify the determinants of a factor of interest (in this case: economic growth). Because the methods discussed all use automatic computer algorithms to do so, they will be referred to as Automatic Search Algorithms (ASA), or Specification Search Algorithms, by which the same is meant.

Methods that make heavy use of statistical information for the specification of models are often associated with "data mining", which is generally considered a derogatory term in economics. This is odd, as this concept has recently become a very popular subfield in the computer sciences. One reason for this is that in economics, datasets are generally quite small, which makes it harder to retrieve reliable information from the data. An important note with respect to data mining as a derogatory term is that data mining with the purpose of finding statistically attractive results is very different from methods that use theoretical justification to retrieve reliable information from data sets. Therefore, the former can be called strong data mining and the latter weak data mining (from Hendry and Krolzig, 2004). This distinction is important in order to be able to address issues with respect to the reliability of weak data mining methods in a neutral way.

General-to-specific modeling is one method to select models automatically (e.g. Hoover and Perez, 1999; 2000; Hendry and Krolzig, 1999; 2003; 2004). Prediction based methods are another (e.g. RETINA, see White, 2000 and Perez-Ameral, Gallow and White, 2003). Many Bayesian approaches to deal with this issue exist too. For instance, sensitivity analysis or derivative methods (Leamer, 1983; 1985; Levine and Renelt, 1992; Sala-i-Martin, 1997). Most prominent is Bayesian Model Averaging (henceforth BMA; e.g. Fernandez et al. 2001b; Sala-i-Martin et al., 2004; and Cuaresma and Doppelhofer, 2007; Magnus et al., 2010 is BMA inspired alternative). While all these methods provide some theoretical justification for their approaches, many have been received with skepticism (see Hoover and Perez, 1999; 2000; 2004; Du Plessis, 2009). What exactly we can learn from these methods is not exactly clear from the theoretical discussions alone. It is often unclear to what extent these methods are able to provide reliable knowledge. Moreover a unified approach is

missing just as much as it is if it comes to growth theory. In order to provide some clarity, a number of studies have assessed the efficacy of these methods with regards to inference (e.g. Hendry and Krolzig, 2003; Castle et al., 2011), or compared different methods of model selection (e.g. Perez-Amaral, Gallo, White, 2005; Magnus, Powell, Prufer, 2010; Hendry and Krolzig, 2004). However, most approaches have focused on the efficacy of these methods in general, without taking into account the specific context of growth economics. This is odd, as a main motivation for this line of research has been growth economics from the start.

An exception to this tendency is Hoover and Perez (2004). They consider model selection methods used in three different seminal papers - of which two were applications to growth economics - (Levine and Renelt, 1992; Sala-i-Martin, 1997; and Hoover and Perez, 1999) in order to assess how reliable these methods really are. In order to do so they design a Monte Carlo experiment. Using data from Levine and Renelt and Sala-i-Martin respectively, they simulate variables replicating growth variables based on a model designed by the authors. They then assess the efficacy of the methods based on their ability to select the "true causes" of the simulated variables. Firstly, it is important that the set of variables that are identified does not include many variables that are no true causes at all. This is captured by the concept of *size*: the probability of falsely rejecting the hypothesis that a variable is not a true determinant. Secondly, it is also important that the variables that are true causes are part of the set of selected variables. This is captured by concept *power,* or, *potency*: the probability of correctly identifying a true cause. They conclude that while Levine and Renelt's sensitivity analysis was too strict (it did not select variables that were in fact true causes), and Sala-i-Martin's model averaging was too lenient (selecting too many regressors as robust), Hoover and Perez' own general-to-specific algorithm (henceforth *Gets*) was "just right" (p.32), achieving a size of just over 5%, while retaining high levels of power.

This assessment provided an important insight in the usefulness of ASA's for solving the growth economics puzzle. However, there are a number of potential problems with arguments such as those made by Hoover and Perez (2004). There are many important differences between a simulation experiment and the real world. The real world often tends to be much more complex than the linear models we would like to fit it on. If we truly want to know how ASA's can help us with puzzles such as the cross-country growth question, we need to learn more about them than Hoover and Perez' experiment provides. Three differences between Hoover and Perez' simulation experiment and the real world as it might be are particularly important for the evaluation of current ASA's. Firstly, much theoretical work has stressed the possibility of nonlinearities and heterogeneity in the economic growth of countries as real possibilities (e.g. Durlauf and Johnson, 1995; Durlauf, Kourtellos, and Minkin, 2001; Kalaitzidakis et al., 2001; Masanjala and Papageorgiou, 2004). Hoover and Perez (2004) however, only consider simple linear models to be truly causing their simulated growth data. Recent work in many branches of model averaging has worked on incorporating heterogeneities or nonlinearities (*e.g.* Cuaresma and Doppelhofer; 2007; Salimans, 2012 (for BMA); Castle and Hendry, 2012 (for *Gets*)). Secondly, in the experiment that Hoover and Perez (2004) designed, only 34 potential regressors were considered. Most research papers on this topic consider a much larger number (in particular Durlauf et al., 2005). The number of potential regressors may greatly affect how well a certain specification search algorithm works. Thirdly, the two Bayesian methods Hoover and Perez assessed have become outdated. Newer methods have been developed and gained much more popularity than the ones Hoover and Perez used. The older methods are unlikely to say much about the efficacy of the newer methods.

In earlier work (Van der Deijl, 2013), the conceptual issues in the debate about automatic model selection mechanisms and data mining methodologies in economic practice has been assessed. In this thesis I ask: how can ASA's help us solve the question why some countries are rich and others poor? A simulation experiment is presented that is similar to Hoover and Perez' (2004) experiment, using up-to-date versions of the model averaging approaches, intended to test the ability of these methods to deal with larger sets of potential regressors, heterogeneous growth patterns and nonlinear relations. This is done in order to get a better picture of how these methods may help us with the long and difficult puzzle of finding the causes of growth.

First, an overview of growth economic theory and empirics is provided, and it is explained why model selection methods ended up being important for growth economics. Then the different approaches to model selection are discussed in more detail (section 2), the design for the simulation experiment is discussed (section 3), and the results are provided (section 4). In section 5 the methods are applied to the real growth data; using the knowledge gained from earlier sections. In section 6 a discussion is provided on methodological and theoretical consequences for the growth economics debate. Section 7 summarizes and concludes.

2. Motivation and Literature Review

2.1 Growth Theory and Open-Endedness[1]

Until the late 1980's, the main way to understand growth, theoretically, was by means of the neoclassical growth models. A seminal contribution in this regard is Robert's Solow growth model (1956). Others include the Ramsey-Cas-Koopmans model and the Diamond's overlapping generations model (see Romer, 2011, for a textbook account). These models contained a small number of factors and were analytically tractable. In its simplest form, the Solow model was a simple expression of the economy as a Cobb-Douglass production function (1), a population growth function (2), and a saving function that links current production to future capital investment given a depreciation factor (3):

$$Y = AK^{\alpha}L^{\beta} \tag{1}$$

$$L_{t+1} = L_t(n + 1) \tag{2}$$

$$K_{t+1} = K_t + sY - dY \tag{3}$$

where Y is total output, A a parameter that reflects technological advancement, K the total amount of available capital, and α its corresponding effectiveness parameter, L the total amount of available labor in society, and β its corresponding effectiveness parameter. Furthermore, n represents population growth, s is a saving parameter, and d is a depreciation parameter. The steady state is reached when the amount of saving is offset by the total depreciation. In this case, investment merely sustains levels of production, but does not increase them. Output growth, in this stage, is merely determined by technological progress (which is an exogenously given factor) and population growth. Extensions to the model, namely the Ramsey-Cas-Koopmans model (Cass, 1965; Koopmans, 1965) or Diamond's overlapping generations model (Diamond, 1965) tent to endogenize the savings function by means of a utility maximization process. However, in the steady state, the exogenously given technological process and population growth remain the sole determining factors of economic growth.

While the tractability, simplicity and intuitive strength of the neoclassical growth models ensured its orthodoxy for three decades, there were two main disadvantages that inspired a new class of models in the late 1980's under the heading of Endogenous Growth Theory (henceforth EGT; Romer, 1994). Firstly, there was a great dissatisfaction among growth economists about the fact that the neoclassical growth models sketched a causal mechanism by which an economy would inspire investments up to the point of a steady-state, but left the causes of growth thereafter completely exogenous. In economic reality, we suppose that there are many causes of economic growth that appear to be caused themselves by the state of the economy (think of educational and factors and economic institutions, such as the state of financial markets). The neoclassical growth models do not consider such factors at all, and this seems highly unsatisfactory. Secondly, Solow's model did not say anything about why differences in growth patterns between different countries would emerge. It was considered an implication of the neoclassical growth model that we would expect to observe a conversion of economies in the world, because of the similar way different countries are treated, and growth was expected to decrease over time[2]. In poorer countries, there is

[1] As I have written on this topic before (Van der Deijl, 2013), some overlap is to be expected in this narrative (in particular with chapter 2.2 in Van der Deijl, 2013).

[2] It is important to note that Solow (1956) does not explicitly assume that the growth curve is similar for all countries. Hence, the convergence is not necessarily assumed in the model (see Chatterji, 1992; Durlauf et al., 2001).

more production advantage to be gained due to investment, and richer countries are likely to be closer to the steady state in which investment is unlikely to result in high levels of economic growth. However, this is not what was observed in the cross-country growth data that was examined, where rich countries turned out to be the ones in which both investment and economic growth were high on average.

Seminal contributions to EGT were Romer (1986) and Lucas (1988). Romer (1986), for instance, endogenizes technological growth by considering research and development practices as part of the economic process. Lucas (1988) considers the possibility of stable returns to human capital. This results in a model without a steady-state, in which investment in human capital keeps its value for economic growth even when countries are already highly developed.

Given the problems the neoclassical growth models suffered from, the EGT models were popular. However, there was also a great disadvantage to the EGT. Whereas the neoclassical growth theories were simple and tractable, EGT opened up the field for a large variety of possible causal factors that could explain growth. Brock and Durlauf (2001) have aptly characterized this field of research as "open-ended": it has been aimed at identifying the many possible causes of growth, and has not aimed at reducing the set of possible causes. Many models propose different causal factors of growth without excluding others. Durlauf et al. (2005), for instance, identify 43 distinct growth theories that propose over 145 different explanatory factors that are considered important for growth. This does not only make the application of "economic theory" theory in reality particularly challenging, it also causes a major problem for the empirical testing of these theories.

Empirical challenges

The growth economics project is ultimately aimed at identifying the true structure of the following regression model[3]:

$$y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \varepsilon \tag{4}$$

for countries i = 1, 2, 3....., n, where Y is long-term economic growth, $\boldsymbol{X_i}$ is a vector containing the explanatory variables of long-term economic growth (and a constant), and $\boldsymbol{\beta}$ a vector containing the corresponding parameters; ε is an error term.

In light of the theoretical ambiguity described above, growth economists have turned to growth empirics[4]. The large amount of proposed causes of growth do not provide clear candidate models to test. As discussed, there are a large set of potential variables in $\boldsymbol{X_i}$ which leads to a large amount of models. Hendry and Krolzig (2004) find that if 41 variables are considered, there are over 2 trillion possible models (2^{41}), whereas a billion billion (2^{62}) possible models are found when 62 variables are considered. If we use Durlauf et al.'s (2005) number (145), the total number of possible models contains over twice as many zeros. It is easy to see that classical hypothesis testing - which at most allows a researcher to reject a certain model but does not allow a comparison between non-rejected models - will not be of much guidance in finding the true model if there are so many models to be tested. For all we know, all models may provide joint explanatory power (F<.001), and if not all, surely a large amount of models will. It is generally asserted that in case more than 1 model is tested on the data, either researchers should correct their results for multiple

[3]I should add that this is an ambiguous formulation of the project. While there appears to be general agreement that the project at stake is finding the best model explaining growth, Leamer (1978) rejects the notion of a true underlying data generating process, whereas Hoover and Perez (2000; 2004) defend the existence of such a structure. Sala-i-Martin simply uses "true model" is scare quotes. I will simply take over this formulation and assume that there is a true structure, that may or may not be stable, and that there are models that approximate this structure. Finding the best one is the project at stake.

[4]I should add that not all empirical growth economics is addressing this question. A large number of growth economists have turned to the randomization movement (see for instance Angrist and Pischke, 2010 for a defense; Deaton, 2010, for a critical account), which focuses its attention on randomized trial experiments, and if not available, the next best thing: natural experiments, or instrumental variable approaches. These methods are not aimed at finding a true model, but are rather aimed at providing knowledge on specific causal relationships.

testing or the results lose their validity (e.g. Mayer, 2000; Kennedry, 2002; Hollanders, 2011). Jeffrey Wooldridge, in his famous text book, puts this quite boldly: "The results (…) we derived for hypothesis testing, assume that we observe a sample following the population model and we estimate that model *once*." (2009, p.678, my emphasis). It is not difficult to see that this leads to severe difficulties for standard hypothesis testing in case of trillions of possible models that theory provides.

The large amount of possible models becomes more problematic even when two further problems are considered. The first one is that the data availability is severely limited (cf. Ulasan, 2011). Growth economists dealing with the questions discussed above are typically not interested in the question what causes growth next year, but what causes growth in the long run. The long-run is typically taken to be at least 25 years. Annual data was started to be gathered in a systematic way around 1960, and the most common data sets used, use data until the year 1996 (e.g. the often used SDM data set, from Sala-i-Martin et al., 2004). This means that there is only 1 data point available per year. Moreover, data is often only available for a subset of countries, which means that there are only around 100 cross-sectional data points available. This is not a lot, considering the large amount of proposed models. Secondly, there is the problem that many variables used in the cross-country growth research are not mutually independent of one another. In other words, there is much multicollinearity in the cross-sectional growth data sets. This problem is nicely worded by Sala-i-Martin (1997): "If one starts running regressions combining the various variables, variable x1 will soon be found to be significant when the regression includes variables x2 and x3, but it becomes non-significant when x4 is included." (p. 178). One such example is the dependency between the East Asia dummy and the variable that captures the part of the population that is Confucian. Both are strongly related to growth, but only if the other is not included. This makes it difficult to determine whether religious practices or other regional factors affect economic growth.

2.2. Automatic Specification Methods

As a result of the challenges that exist in the field that deals with the evaluation of determinants of growth with respect to cross-country growth data, a number of authors have proposed to use automatic model selection techniques. Edward Leamer was seminal for a movement that wanted to nudge economics towards these methods rather than standard hypothesis testing (1978; 1983; 1985). One of his suggestions (proposed explicitly in Leamer, 1983) is that we should focus on the *robustness* of variables. Key is the idea that if a variable is related to the variable of interest in one model, but not in another (or, even worse, it is aversely related to it), then a variable is *fragile*. Only a variable that is consistently related to growth in a set of alternative models is to be trusted. This approach is called Extreme Bounds Analysis (hencefoth EBA). It was applied in the field of growth economics by Levine and Renelt (1992): a seminal paper for model selection methods in growth economics. However, they cannot find a variable that is robustly related to growth (though investment comes close), and end on a disappointed note. This line of research was picked up by Sala-i-Martin (1997), who argued that Leamer's EBA was too stringent and proposed that variables do not necessarily need to consistently robust in order to be considered important, but they do need to be robust given a weighted average of the parameter estimates throughout the range of considered models. Sala-i-Martin's paper (provocatively titled "I just ran two Million Regressions") was seminal for model averaging approaches in economics. This line of research was further developed by Ley and Steel (Ley and Steel, 1999; Fernandez et al., 2001b) in a fully Bayesian framework, making use of the already available literature on Bayesian model averaging (such as Hoeting et al., 1999). In the meanwhile, alternative approaches also deemed up, and were used in the growth literature. In particular the general-to-specific approach based on the LSE methodology of David Hendry (2000), first developed by Hoover and Perez (1999), took part in the debate (Hoover and Perez, 2004; Hendry and Krolzig, 2004).

Regardless of Hoover and Perez' (2004) positive conclusions with respect to the efficacy of their method vis-à-vis the BMA and EBA, *Gets* has not been used much in the growth economics

debate of the last decade, and the discussion on *Gets* turned into a separate literature. These two lines of research, BMA and *Gets*, both aim at finding correct specifications in a data-driven way. While occasional discussions of each other's work occur (e.g. Hoover and Perez, 2004; Hendry and Krolzig, 2004; Eicher et al., 2007; Castle et al. 2010; Magnus et al., 2010), debates are uncommon. *Gets* is often criticized for not taking seriously the dangers of pretest bias (e.g. Magnus et al., 2010), while Bayesian approaches are criticized for not taking seriously the realism of the data generating structure and the ability of researchers to discover it, and for embracing subjectivity (e.g. Hoover and Perez, 2000; Du Plessis, 2009). In the following the methods are discussed in more detail, and a distinction between the different specific available methods is made within those approaches.

Model Averaging

The first systematic way to deal with model selection in a Bayesian way was to Leamer's extreme bounds analysis (1983; 1985; applied in Levine and Renelt, 1992), which simply existed out of estimated a large set of possible models, and analyze for each parameter whether the variable was significant and had the same sign in all possible models. The extreme bounds for a specific variable of interest thus defined as follows[5]:

$$[(\min_i(\widehat{\beta_i} - 1.96(\hat{\sigma}(\widehat{\beta_i}))); \max_j(\widehat{\beta_j} + 1.96 * (\hat{\sigma}(\widehat{\beta_j})))] \quad (5)$$

where, $\hat{\sigma}$ is the standard error of estimation, i = 1, 2, 3, …. m, j = 1, 2, 3,… m are model indexes, and m is the total set of models.

While not yet fully developed, the first moves towards a systematic approach to model averaging in the growth literature was Sala-i-Martin's (1997) approach. Sala-i-Martin proposed that EBA robustness checking was right in spirit, but too stringent. In order to avoid this, Sala-i-Martin proposed to average the parameters and corresponding standard error statistics by means of a weight determined by the likelihood of the estimated models.

$$\hat{\beta} = \sum_{z=1}^{M} w_z \widehat{\beta_z} \quad (6)$$

For z = 1, 2, 3…M models. The weights are a relative likelihood function of the specific model to the rest:

$$w_z = {L_z}/{\sum_{z=1}^{M} L_z} \quad (7)$$

Similarly, the variance of these estimates is taken to be a weighted average too:

$$\hat{\sigma}^2 = \sum_{z=1}^{M} w_z \hat{\sigma}_z^2 \quad (8)$$

The bounds are defined as [$\hat{\beta}$ -1.96$\hat{\sigma}^2$; $\hat{\beta}$ +1.96$\hat{\sigma}^2$][6]. Whenever this bound does not contain zero, it is robust.

While Sala-i-Martin's method and the EBA laid the foundations for the Bayesian model averaging approaches, the field quickly developed further. These two methods are instrumental in

[5] This formulation is borrowed in an adjusted form from Sala-i-Martin (1997)

[6] That is, if the variable is normal. If not, the procedure is slightly more complex (see Sala-i-Martin, 1997).

understanding the developments, but unlike Hoover and Perez (2004), they will not be assessed in the experiments presented here. Both methods are attractive in their comprehensibility, but lack theoretical justification (acknowledged, for instance, in Sala-i-Martin et al., 2004). Moreover, Sala-i-Martin's method is particularly restrictive, as it is designed only to examine models that are of the same size in terms of the number of included variables. Instead, further developed versions in the same spirit are assessed: BMA (first applied to growth economics by Fernandez et al., 2001b) and Bayesian Averaging of Classical Statistics (henceforth BACE; first developed by Sala-i-Martin et al., 2004).

Bayesian model averaging

Fernandez et al., (2001b) laid the groundwork for the application of this method to the problem of open-endedness in growth economics. Moral-Benito (2012) and Hoeting et al. (1999) provide a clear summary of the BMA methodology. The key rationale behind BMA is that statistical methods may not be able to determine which model is the true one (or which ones are to be rejected), but is able to attributive a certain level of confidence to the likelihood that a particular variable is truly related to growth. In the end, the method is aimed at estimating Pr(β|D): the posterior density function of the values β can take, having observed the data, where β is the parameter linking a variable of interest to growth, and D is the data. The information for estimating this probability is derived from the estimation of the parameters in different models, where M_z refers to a specific model z . This posterior density, using models, can be written as:

$$Pr(\beta \mid D) = \sum_{z=1}^{M} Pr(\beta \mid M_z, D) Pr(M_z \mid D) \tag{8}$$

Where ($\beta \mid M_z, D$) is the posterior probability density function of model parameters given model z and the data. $Pr(M_z \mid D)$ is the posterior probability of model z, such that $\sum_{z=1}^{M} Pr(M_z \mid D) = 1$. This latter assumption boils down to saying that the true model is contained in the considered models; an assumption that can be relaxed (see Geweke, 2010). In order to get at $Pr(\beta \mid D)$, further specification of the terms pr(β | M_z,D) and (pr(M_z | D) is required. Both can be reformulated by means of Bayes' rule (or the law of total probability). Firstly, using Bayes' theorem, we can write the pr(β | D, M_z) as

$$Pr(\beta \mid D, M_z) = \frac{Pr(D|\beta, M_z) Pr(\beta|M_z)}{Pr(D|M_z)} \tag{9}$$

Furthermore the model probability given the data can be rewritten as:

$$Pr(M_z \mid D) = \frac{Pr(D|M_z) Pr(M_z)}{Pr(D)} \tag{10}$$

We can use (9), integrate it to find a further specification of pr(D | M_z) in (10).

$$pr(D \mid M_z) = \int pr(D \mid \beta, M_z) pr(\beta \mid M_z) d\beta \tag{11}$$

Finally, the aim of most applications of Bayesian model averaging is to find posterior inclusion probabilities, Pr(β ≠ 0 | D), of the variables included. This is found by using the expected value and variance of β:

$$\text{E}(\beta \mid \text{D}) = \sum_{z=1}^{M} \hat{\beta}_z \, pr(M_z|D) \tag{12}$$

$$\text{Var}(\beta \mid \text{D}) = \sum_{z=1}^{M} (Var[\beta|D, M_z] + \beta_z^2) \, pr(M_z|D) - [\text{E}(\beta \mid \text{D})]^2 \tag{13}$$

From (8) - (11) we can see that the posterior, Pr(β | D), can be specified in terms of the pr(D | β, M_z), pr (D | M_z), pr(β | M_z) and pr(M_z). The latter two are prior probabilities that are data independent[7]. pr(M_z) is the model prior, and pr(β | M_z) is called the parameter prior, or prior on parameter space. The problem with specifying priors is that there are so many models and accessory parameters estimated, and identifying our prior beliefs for each of these parameters would be infeasible. Particular challenges are that we would like to represent ignorance about our prior knowledge of the parameters, as we generally know very little about them. Representing our very limited knowledge mathematically, while at the same time keeping computation simple, turns out to be quite complex (see Norton, 2010, for clear conceptual discussion on this issue; Eicher et al., 2011, for a technical assessment). The same applies to some extent to priors on model space. While generally a priori knowledge does not allow us to favor any one of the M considered models over any other considered model, often we might want to give more priority to simpler models, or models whose size comes closer to our expectation (see Sala-i-Martin et al., 2004; Cuaresma and Doppelhofer, 2007). This is not altogether uncontroversial though (see Magnus et al., 2010).

A large variety of different methods exist to arrive at parameter priors. Eicher et al., 2011, for instance, study 12 different specifications. A popular way of modeling prior knowledge is by means of Zellner's g-prior (from Zellner, 1986). G-priors are defined such that the variance is a function of a parameter σ^2 and $X_z'X_z$ (depending on the model): $g(\sigma^2 X_z'X_z)$, where g is a scalar function (given g-priors its name). There are a number of alternative options for the selection of this parameter. Fernandez et al. (2001a) select this scalar to depend on both the number of models used (q) and the sample size (N): g = max (N,q), which they claim to have the most desirable predictive features. They call these the benchmark priors. Steel and Ley (2009) and Eicher et al. (2011) summarize different ways in which this can be done and analyze the efficacy of these choices with respect to the context of long-term growth empirics.

While using g-priors is common, it is also somewhat controversial as it uses the data to specify the functional form of the prior, which does not conceptually fit very well with the Bayesian philosophy. Some methods of parameter prior specification do not use such information. Under certain formulations of such parameter priors that represent ignorance, such as chosen by Sala-i-Martin et al. (2004) the expected value of (8) can be written as follows (Moral-Benito, 2012):

$$E(\beta \mid D) = \sum_z^M Pr(M_z \mid D)\hat{\beta} \tag{14}$$

where $\hat{\beta}$ is the classical maximum likelihood estimator of β. This is the reason why Sala-i-Martin et al. (2004) call their method the Bayesian Averaging of Classical Estimates.

Model priors are sometimes dependent on the number of regressors included in the model and the prior probability of each variable included in the model, which is often modeled by means of a binomial distribution (Moral-Benito, 2012). If all models, with the same size, are considered a priori equally likely, this distribution simply collapses into $P(M_z) = 2^{-q}$ (used by Fernandez et al., 2001b). Sala-i-Martin (2004) choose and alternative method. It requires a specification of a prior hyper-parameter "expected model size" q̂. Rather than specifying prior model probabilities, they specify prior variable inclusion probabilities, which are defined as q̂/Q, where Q is the total number of regressors.

A final note to be made is that most Bayesian model averaging techniques requires so much computation that it is infeasible to do, even considering the developments in computational power. Generally a subset of regressions is run by means of Markov Chain Monte Carlo (MCMC). This means that only a subset of possible models is visited, and the model averaging occurs on this subset of models. For instance, in the application of BMA that is used in this paper, a birth-death sampler is used (Zeugner, 2012). Given an already selected model, a new model is selected by adding a new randomly chosen covariate. If this covariate is already part of the existing model, it is

[7] As will be discussed shortly, priors can be made data-dependent, but formally, they need not be, and conceptually, they represent the belief of the researcher prior to having viewed the data.

dropped from the model. If it was not, it is added. In this way a subset of models is visited.

Weighted-average least squares (WALS)

While in the BMA tradition, Magnus et al. (2010) provides an alternative approach to standard BMA approaches. Confronted with the problems that model averaging techniques are generally highly susceptible to multicollinearity problems, and are very heavy on computer power even if random sampling methods like MCMC are used, Magnus et al. (2010) proposes very different method to get to the estimates in a Bayesian averaging framework. The method departs with the observation that many of the reasons why we would want to average so many different regressions is that growth regressors are not orthogonal, i.e. they are not independent, but highly collinear. Orthogonalizing regressors would simplify the process of estimating the relevant parameters enormously. In fact, in case all potential regressors would be orthogonal, all the estimated parameters would be the same in all possible models. In comparing their approach to BMA, they conclude that the main difference in estimation lies in estimating the prior on parameter space. Magnus et al. use the La Place distribution to model these priors. Additionally, they argue that the tendency to make model priors size dependent wrongly biases the estimation towards parsimonious, small, models. Hence, a flat prior over model size is used in their approach. A useful advantage of the approach is that this greatly reduces the set of necessary models to be estimated to assess the relationship between the set of regressors and the dependent variable.

The general-to-specific approach

Gets was developed from the LSE methodology to econometrics and its theory of reduction developed by David Hendry and co-authors (see Hendry, 2000, for a collection of his work on this matter; Campos et al., 2005, for a summary and overview). The LSE methodology is based on the assumption of a true data generating process (DGP), meaning an ideal econometric model maps real economic processes. Furthermore, it is based on the insight that a large general model (General Unrestricted Model; henceforth GUM) should be able to capture the main structure of the data. Whether a model is able to do this is not only dependent on model fit, but also on a large number of additional features, such as the absence of structural breaks, absence of serial correlation, absence of heteroskedasticity and no signs of incorrect functional form. These features are jointly called *congruence*. In order to be congruent, a model needs to be theory consistent as well. If a model is able to meet all the requirements, it is said to be congruent, and it can be said to capture the main structure of the data. If this is not the case, it is unlikely that a model nested in the GUM would be congruent. In case a GUM is congruent, all the unnecessary features of the model are removed while ensuring that the reduced versions of the model still captures the main structure of the data. The aim of the method is to find the smallest model that explains all the features of reality that were captured by the large model. Such a model is *encompassed* in the GUM. This model is said to describe the local data generating process (LDGP).

The general-to-specific algorithm was first developed by Hoover and Perez (1999), and further developed by Hendry and Krolzig (1999; 2003) and many subsequent papers of Hendry and collaborators. More recently a augmented version of the algorithm has been implemented as a OxMetrics search algorithm: *autometrics* (Doornik, 2009). The algorithms are basically automated realizations of the methodology. It starts with a GUM that contains all potential regressors. Then, this GUM is tested for congruence by means of diagnostic testing and tests of statistical fit. Insignificant parameters are removed, after which the model is checked again for congruence. If a removal leads to a violation of the congruence, it is retained. This is done until a final model is reached that cannot be reduced. If multiple variables can be removed from a model in a congruent way, leading to alternative models, the final models are compared by means of the Bayesian information criterion, after which the best model is chosen.

A note on methodology needs to be made with respect to the multiple testing that is involved in this procedure. Skepticism (like expressed by Keuzenkamp, 1995; Magnus et al. 2010) about the method is based on the fact that it runs multiple hypothesis tests (t-tests) to select regressors, which

may lead to a large pre-test bias. A number of counter arguments have been phrased against these worries. Firstly, Hendry (2002; Hendry and Krolzig, 2003) stresses that the costs of search that this multiple testing causes is relatively low. The chance of making errors is only slightly increased, and can even be contained by using more stringent tests. One reason is that the probability of retaining false (but significant) coefficients is still very small if the critical t-value is selected to be small enough. A second defense is more radical. Hoover and Perez (2000) simply claim that their method is not based on the classical inference pattern of frequentist statistics, but is rather based on the notion of congruence in which t-values are not interpreted inferentially, but in which a high t-value is simply an indication of good fit. The most convincing defense perhaps, is simply pragmatic. In Monte Carlo simulations that have been conducted by proponents (e.g. Hoover and Perez, 1999; 2004; Hendry and Krolzig, 2003) *Gets* turns out to retain high levels of power while often retaining a size smaller than, or close to, 5%. In the least, this shows, that the pretest bias worry that some share in relation to the *Gets* methodology is not necessarily founded.

Nonlinearities and heterogeneities

While these approaches (in particular the BMA) already consider a large set of potential models, the number of models is still very limited compared to the total set of models that can be made with the same set of variables. One particular restriction that has worried many is the exclusion of nonlinear models from the set of considered models. There is quite some support in the theoretical literature on growth that nonlinear models are at least possible, and there is in fact evidence that they are quite likely there (e.g. Durlauf and Johnson, 1995; Papageorgiou, 2002).

Both in the *Gets* literature as well as the BMA literature attempts have been made to develop selection algorithms that can detect nonlinearities. In terms of nonlinearities, Cuaresma and Doppelhofer (2007) and Salimans (2012) provide extensions to the BMA/BACE framework, whereas Castle and Hendry (2012) provide a way to incorporate nonlinearities in the *Gets* framework.

nonlinearities in a Bayesian framework

Cuaresma and Doppelhofer (2007) use a threshold approach. This means that a function is estimated with a number of potential nonlinearities that are included if their posterior inclusion probability exceeds a certain threshold. The threshold regression that they consider looks as follows:

$$y = a + \sum_{k=1}^{n} x_k \beta_k + \sum_{j=1}^{m} [(a_j^* + \sum_{k=1}^{n} x_k \beta_k) l(z_j < t_j)] + \varepsilon \quad (15)$$

Where z_j is an indicator variable and l is an indicator function that takes on the value 1 if the statement is true and is zero otherwise.

Their methodology follows Sala-i-Martin et al. (2004) closely. It estimates and averages linear models existing out of a subset of the total set of linear regressors, X, but now allows for the threshold nonlinearities of these variables with variables in a given set of potential nonlinear variables, Z. The thresholds correspond to prior inclusion probabilities. If, for a single observation, the threshold is exceeded, the additional parameter is added on to the expected growth. While computationally heavy, the method is a conceptually clear elaboration on the Bayesian model averaging framework.

Nonlinearities in a Gets framework

Considering nonlinear terms as potential regressors, means a large expansion of the set of potential variables. A key feature of Doornik's general-to-specific algorithm *autometrics* (2009) is an algorithm that allows to search among regressors when the set of potential regressors is larger

than the number of observations. This feature makes it computationally feasible to add larger sets of regressors to a GUM, while retaining the ability to reduce it to a more parsimonious model. This opens the door for dealing with nonlinearities in a general-to-specific framework. This approach to functional form uncertainty has been taken by Castle and Hendry (2012). The key idea is that if the true DGP is nonlinear, the correct nonlinear functions in the GUM are more likely to be retained than its linear counterparts. By simply adding all potential functional forms, the general-to-specific algorithm can simply select from them the best one. Some restriction exist though, square roots of regressors, for instance, are not possible if the regressor contains (many) negative values. Castle and Hendry, propose to apply a Taylor expansion, resulting in the following GUM for K potentially important variables:

$$\mathrm{y} = a + \sum_{k=1}^{n} x_k \beta_k + \sum_{i=i}^{k} \sum_{k=1}^{n} x_k x_i \beta_{i,k} + \sum_{j}^{i} \sum_{i=i}^{k} \sum_{k=1}^{n} x_j x_k x_i \beta_{i,k} + \varepsilon \tag{16}$$

However, there are a number of difficulties, and potential solutions described in Castle and Hendry will be discussed briefly.

Firstly, entering this amount of functional forms, quickly makes the GUM very large. Hendy and Caslte argue that this is a threat in particular to making false inferences, because even if the size of the selected test is small, making many inferences will increase the chances of making false inferences. Therefore, they propose as a solution to the problem to select a very small p-value for the automated strategy.

A second problem is that variables will be included now in many different functional forms, and these functional forms are likely to be related. For instance, the squares of a variable are likely to be correlated with the levels. And, if a variable was already strongly correlated with another, their interaction will be correlated strongly to the squares of the original variables. Hence, this approach will likely result in high levels of multicollinearity. Hendry and Castle argue that this problem can largely be solved by double de-meaning. Demeaning both the level variable and the squared variable, will decrease their strong correlation, and make it easier for the program to detect which variable can explain which part of the variance in the dependent variable.

Lastly, a problem with functional form detection is that they are very sensitive to outliers. One strong outlier can make an originally linear relation appear to have an alternative functional form. In order to avoid this, Hendry and Castle propose a way to detect outliers to deal with this issue which is called impulse-indicator saturation (IIS; discussed in Santos et al., 2008). The basic idea is that in the general-to-specific algorithm dummy's for every specific observation are added to the GUM, which are removed if the observations can be better explained by the other variables in the GUM. This turns (16) into:

$$\mathrm{y_p} = a + \sum_{k=1}^{n} x_k \beta_k + \sum_{i=i}^{k} \sum_{k=1}^{n} x_k x_i \beta_{i,k} + \sum_{j=1}^{i} \sum_{i=i}^{k} \sum_{k=1}^{n} x_j x_k x_i \beta_{i,k} + \sum_{k=1}^{n} \delta_k 1_{(k=p)} + \varepsilon \tag{17}$$

Where $1_{(k=p)}$ is an indicator for the p^{th} observation. This method ensures that the variance detected by the nonlinear specification algorithm is not just based on outliers, but that real nonlinearities are detected.

3. The experiment

Motivation

There are many challenges in learning how well search algorithms such as the ones discussed here are likely to perform if applied to the cross-country growth. Ideally, we would like to learn how high the probability is that methods select the correct variables. A way to do this in general cases is to assess the predictive performance of the models generated by the different methods. However, given the scarcity of the long-term growth data, this is not quite feasible. An alternative method is taken by Hoover and Perez (2004). They propose to simulate data that are as similar to real growth data as possible. While in case of real growth data, we still have to find out what truly causes growth, in the simulated setting we can construct our own DGP. If the methods are successful in retrieving these structures, we can be confident that they will do so too in reality. This is only sound though, if the simulated DGP's and the DGP behind long-term growth are sufficiently similar. It is therefore crucial in this approach to make realistic assumptions about all the aspects of the DGP that matter to inference.

Hoover and Perez (2004) applied this testing method and concluded that their *Gets* algorithm did particularly well in uncovering the DGP vis-à-vis Levine and Renelt's (1992) method, and Sala-i-iartin's (1997) method, and performed very well too in relation to classical methods with a priori knowledge of the DGP. However, there are three problems with Hoover and Perez (2004) experiment that motivate this project.

Firstly, in the past decade much has happened and newer and better method are available, that have been described above. It is important to see if the newer methods may do better than the ones that Hoover and Perez assessed.

A second concern with Hoover and Perez (2004) is that growth was modeled in a very simple way. Many growth economists have argued that the kind of relations that govern the long-run economic growth variation are not as simple and linear as the earlier growth economists considered (cf. Durlauf and johnson, 1995; Durlauf et al., 2001; Masanjala and Papageourgiou, 2004). Jeffrey Sachs's well-known concept of a "poverty trap" is a simple exemplification of a hypothesis about a nonlinear process behind long-term economic growth (Sachs et al., 2004).

Lastly, a related worry is that the experiment conducted by Hoover and Perez (2004) used a set of 34 variables, whereas the total set of variables considered by most papers (e.g. Sala-i-Martin, 1997; Sala-i-Martin, 2004; Ley and Steel, 2009; Salimans, 2012) exceeds 60. It is easy to imagine that the task to select the right variables from a set of 34 variables is easier than to do this from a set of 67 variables (like in case of the Sala-i-Martin et al., 2004, dataset). The aim of this study is to investigate to what extend the efficacy of the search methodologies is contingent on the amount of included variables.

In general, the experiment is designed to test the robustness of these specification algorithms to the context in which they are used.

Data

The main source of the data used for this experiment is the commonly used SDM dataset (Sala-i-Martin et al., 2004), which contains 67 variables potential determinants of growth, and a growth variable that is an average over the years 1960-1996. A number of updates were made for this dataset. While a lot of the data was measured in the beginning of the period (1960's), or is simply independent of time (e.g. former British colony, land-locked, country-size), a number of variables are time dependent. This includes the growth variable. For those variables, the dataset was update with data until 2010; that is, if the data was available. Updates were made for the following variables: Average economic growth, Openness, economic growth, population growth interest rates, and squared interest rates (from the world bank data), and the capitalism index, economic freedom index and civil rights index (from the freedom house). The data is summarized in table A1 in the appendix.

Set-up

Throughout the simulation experiments, Hoover and Perez (2004) are closely followed in terms of the general setup. However, rather than one, three versions of the experiment were conducted. The first experiment is based on "nice data". Here only the 42 most important variables according to Sala-i-Martin et al. (2004) are used. The missing values were imputed in the way that Hoover and Perez (2004) recommend: by means of multiple imputations with a ridge prior to solve the fact that there are insufficient data points to do it without. Like in Hoover and Perez (2004) a set of specifications was chosen that relate independent variables to a simulated "growth variable": a variable that is generated to look like the economic growth variable, but of which the true causes in the data are known by design. These specifications, or models, exist respectively out of 0, 3, 7 or 14 independent variables. For each one of these quantities except 0, 10 model specifications were selected to be in the simulated DGP. In other words, for each of these model sizes (3, 7 and 14), 10 different groups of variables are selected. Such a set of 3, 7 or 14 models is a specification. Then, for each specification, 100 "simulated growth variable" are generated of which we thus know the "true data generating structure". These variables differ only due to their error term, which is randomly sampled from the estimated residuals of estimation. Each set of 100 simulated variables with the same model specification will be referred to as a "dataset". In order to obtain the coefficients that relate the independent variables to the simulated growth variable, the following procedure is followed that closely follows Hoover and Perez:

1. A random selection of 0, 3, 7 or 14 variables, j, from the 42 potential independent variables from the data set was made.
2. The regression $y = \boldsymbol{\beta}\mathbf{X_j} + \mathbf{u}$ is run, and the coefficients of $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are obtained, where y is the real growth variable, and $\mathbf{X_j}$ the set of randomly selected independent variables, $\hat{\boldsymbol{\beta}}$ the estimated parameters on $\mathbf{X_j}$, and $\hat{\mathbf{u}}$ the estimated residual.
3. For each i = 1, 2, 3, 4.....100, a variable y_{ij} is simulated, such that $y_{ij} = \boldsymbol{\beta}\mathbf{X_j} + \hat{\mathbf{u}}_i^*$, where $\hat{\mathbf{u}}_i^*$ is bootstrapped using a wild bootstrap from $\hat{\mathbf{u}}$
4. The search algorithms were run in order to see if they retrieve the correct set of variables that were indeed in $\mathbf{X}_j$.

This procedure provides, in total, 4000 simulated variables in a total of 40 datasets. The significance of this procedure is that of these simulated variables - different from real world applications - we now know the true DGP. In order to evaluate the ASA's, the experiment consists of running the ASA's over the simulated variables in order to see if the true data generating structure is discovered.

The reason a set of different specifications is chosen, rather than a single one is that the strength of the relationship between the regressors and the dependent variables may differ in different specifications. We shall refer to this relationship as the signal-to-noise ratio. It is operationalized by the probability that the true relationships are identified (are significant), when a regression with the true specification is run (see the discussion on "true power" below).

What is important is that while the complete 40 datasets were used for the WALS and *Gets* algorithm (that were relatively time inexpensive to run), only 10% of the simulations were used for the evaluation of BMA, of which each dataset consisted of only 10 simulated dependent variables (rather than 100). This was due to the computational burden of BMA, which takes roughly 10 minutes per run, boiling down 67 hours of computer run time for 400 runs (two and a half day and nights). Running the complete datasets for BMA would have cost over three weeks of running my computer day and night per experiment.

In a second version of the experiment, the same procedure is followed. However, rather than using the nice data, the full updated SDM data set is used, with 67 variables. Moreover, the data set is not imputed in the same way that it is done in the first version of the experiment, meaning that we now have to deal with missing data. Another set of specifications is generated and the same procedure is repeated for these specifications.

Experimenting with nonlinearities

A third Monte Carlo experiment is intended to test the robustness of specification search algorithms to functional form misspecification. In order to do this, a number of adjustments have to be made. Experiment 1 and 2 are intended to assess the search algorithms for robustness to size of the set of potential regressors and the "niceness" of the data. The third experiment is designed to assess how ASA's efficacy reacts to alternative functional forms. In order to do so, three specifications were chosen of size 3, 7 and 14 that consisted of (a subset of) the 14 variables that are identified by either one of the three specification search algorithms as one of the main variables to play an important role in determining the true growth variable. This selection was made in order to ensure that the signal-to-noise ratio is at least high enough to observe a difference. These specifications were then altered by adding nonlinear components to them (see table 1). Then, like in experiment 1 and 2, a number of simulated growth variables created. First, the three specification algorithms that were used in experiment 1 and 2 were evaluated, but, the *Autometrics* algorithm intended for cases in which there is functional form uncertainty discussed in Hendry and Castle (2012) is assessed too. Due to limited computer power, however, it was not possible to include the Cuaresma's and Doppelhofer's BAT approach in the experiment[8].

In table 1, the specifications that were used are summarized in more detail. On the basis of all specifications is the following model:

$$y_{ij} = \boldsymbol{\beta}\mathbf{X_j} + \hat{\mathbf{u}}_i^* \qquad (18)$$

Where, $\mathbf{X_j}$ is either a set either of 3, 7 or 14 of the variables that rolled out of the specification algorithms as the best ones. Keeping these variables the same has the disadvantage that contingencies with respect to these specifications, such as high levels of multicollinearity between two variables, can greatly affect the results. However, it does make it easier to compare the effect of nonlinearities separately from other aspects of specification.

The inspiration for the different variations on the linear functional form were taken from the literature on growth theories. The idea behind this is that if these theories were correct, successful ASA's should be able to do their work and identify the correct specification, or, at least be able to still identify the linear variables that matter in the specification. While the theoretical background of these ideas does not really matter for the experimental design, the existence of these theories inspire the question: if they were true, would the ASA's still work?

Table 1: functional forms for the third experiment

1	$y_{ij}=\boldsymbol{\beta}_a\mathbf{X_j} + \hat{\mathbf{u}}_i^*$ The variables in **X** exist out of 3, 7 and 14 variables selected for doing well in the different specification algorithms[a, b, c]
heterogeneity	
2	If years open >.2 :$y_{ijl}=\boldsymbol{\beta}_l\mathbf{X_j} + \hat{\mathbf{u}}_i^*$ If years open <.2 :$y_{ijm}= \boldsymbol{\beta}_m\mathbf{X_j} + \hat{\mathbf{u}}_i^*$ Where the estimates of $\boldsymbol{\beta}_l$ and $\boldsymbol{\beta}_m$ come from different OLS estimations, such that all the parameters on $\mathbf{X_j}$ are different.
3	If country is african of south american :$y_{ijl}=\boldsymbol{\beta}_l\mathbf{X_j} + \hat{\mathbf{u}}_i^*$ otherwise :$y_{ijm}= \boldsymbol{\beta}_m\mathbf{X_j} + \hat{\mathbf{u}}_i^*$ Where the estimates of $\boldsymbol{\beta}_l$ and $\boldsymbol{\beta}_m$ come from different OLS estimations, such that all the parameters on $\mathbf{X_j}$ are different.

[8]Running the algorithm with sufficient replications takes about 20 hours on my home laptop. The amount of time it would take to run this some 200 times would be too much for this project (roughly 5 months for just this).

4	If a variable belongs to a randomly generated partition :$y_{ijl}=\boldsymbol{\beta_l X_j} + \widehat{\mathbf{u}}_i^*$ otherwise :$y_{ijm}= \boldsymbol{\beta_m X_j} + \widehat{\mathbf{u}}_i^*$ Where the estimates of $\boldsymbol{\beta_l}$ and $\boldsymbol{\beta_m}$ come from different OLS estimations, such that all the parameters on $\mathbf{X_j}$ are different.
5	If the log of GDP in 1960<7.1 :$y_{ijl} = \boldsymbol{\beta_a X_j}$ *1.5 + $\widehat{\mathbf{u}}_i^*$ If the log of GDP in 1960>7.1 :$y_{ijm}= \boldsymbol{\beta_a X_j} + \beta_b$rerd + $\widehat{\mathbf{u}}_i^*$, Where $\boldsymbol{\beta_a}$ is the same as $\boldsymbol{\beta_a}$ in specification 1, rerd is exchange rate distortion, and the β_b a slightly inflated estimate of its effect on growth.
6	If the log of GDP in 1960<7.1 :$y_{ijl} = \boldsymbol{\beta_a X_j}$ *1.5 + $\widehat{\mathbf{u}}_i^*$ If the log of GDP in 1960>7.1 :$y_{ijm}= \boldsymbol{\beta_a X_j} + \beta_c$h60 + $\widehat{\mathbf{u}}_i^*$, Where $\boldsymbol{\beta_a}$ is the same as $\boldsymbol{\beta_a}$ in specification 1, h60 is higher education enrollment in the 1960's, and the β_c a slightly inflated estimate of its effect on growth.
nonlinearities	
7	$y_{ij} = \boldsymbol{\beta_a X_j} + \boldsymbol{\beta_b Z_j} + \widehat{\mathbf{u}}_i^*$, Where $\mathbf{Z_j}$ is a set of either 1, 2 or 3 interaction terms, depending on whether the model size is 3, 7 or 14, that include log of GDP in 1960, $\boldsymbol{\beta_b}$ is the estimated effect on growth. The interaction terms are: 1) gdpch60l*zzprights 2) gdpch60l*opendec1 3) gdpch60l*p60
8	$y_{ij} = \boldsymbol{\beta_a X_j} + \boldsymbol{\beta_b Z_j} + \widehat{\mathbf{u}}_i^*$, Where $\mathbf{Z_j}$ is a set of either 1, 2 or 3 interaction terms, depening on whether the model size is 3, 7 or 14, that include real exchange rate distortions, $\boldsymbol{\beta_b}$ is the estimated effect on growth. The interaction terms are: 1) rerd*priexp70 2) rerd*zzprights 3) rerd*gdpch60l
9	$y_{ij} = \boldsymbol{\beta_a X_j} + \beta_b$Kalait01+ $\widehat{\mathbf{u}}_i^*$, where Kalait01 is a variable inspired by a paper of Kalaitzidakis et al. (2001; figure 2) that used a nonparametric approach to estimate that the effect of growth was nonlinear with respect to initial income: in very low income countries it is negative, thereafter positive, and non-existent for countries that were already developed. The corresponding β_a is the estimated effect of the normal (linear) effect of higher education on growth.
10	$y_{ij} = \boldsymbol{\beta_a X_j} + \beta_b h60^2 + \widehat{\mathbf{u}}_i^*$, where $h60^2$ is simply squared level of higher education enrollment rates and β_a its estimated effect.

a)
- Real exchange rate distortions (1), log GDP in 1960 (2), and investment price (3).

b)
- Real exchange rate distortions (1), log GDP in 1960 (2), and investment price (3),
- primacy schooling in 1960 (4), fraction buddhist in 1960(5), fraction confucian in 1960 (6), malaria prevention in 1966 (7).

c)
- Real exchange rate distortions (1), log GDP in 1960 (2), and investment price (3),
- primacy schooling in 1960 (4), fraction buddhist in 19605), fraction confucian in 1960 (6), malaria prevention in 1966 (7),
- fertility rates in the 1960's (8), years the economy has been open counted from 1994 (9), fraction muslim in 1960 (10), fraction of GDP in mining (11), government consumption share (12), fraction speaking other language (13), population

density in 1960 (14).

The first four nonlinear specifications (2-6) are based on the idea that growth paths of countries may not all be the same for all countries in the world. This idea is often phrased in terms of convergence clubs (originating from Abramovitz, 1986; Baumol, 1986). This is tightly connected to the "twin-peaks" (Quah, 1997) observed in the data, and empirical research on convergence. Whereas we could expect from a universal version of Solow's model that convergence would occur for all countries, the convergence appears to occur in two separate groups, where one of the groups converges to higher levels of income per capita than others. A particularly important development in the literature is the idea that while the Solow model may be correct in identifying that the causes of growth are the same for all countries, the size of the effect they have on growth is not. For instance, Durlauf and co-authors have argued that the Solow model may explain local growth much better than global growth patterns (Durlauf and Johnson, 1995; Durlauf, Kourtellos, Minkin, 2001). Central in all these concept is the idea of heterogeneous growth: the effects of variables that cause growth are not universal for all countries.

In this study, the idea of heterogeneous growth is operationalized by creating a partition in the data set on the basis of split variable. For the two partitions, the causes of growth affect our simulated growth variable differently. This is what happens in specifications 2-6. Specification 2 uses *Years a country has been an open economy* (yrsopen) as a split variable, an idea that is based on Serranito (2004). Specification 3 is based on the idea that variables affect growth differently in different parts of the world. This is an idea that has been played with by (for instance) Masanjala and Papageourgiou (2008), who applied Bayesian model averaging to find out whether countries in Africa have a different growth path than countries in other continents and argue that this is so. In order to facilitate equal comparison, partitions need to be made such that the groups are approximately of equal size. Hence, the cut was made on the basis of whether countries are part of *southern Africa* (safrica) or *Latin America* (laam). Specification 4 randomly partitions two groups of different growth patterns. This is based on a finding in Baştürk et al. (2012) that convergence clubs need not at all follow a logical pattern. In their approach, a data-driven method is used to identify two convergence clubs in Asia, Africa and South America, and find that convergence need not follow a predictable pattern. They find that, just to name a few, Brasil, Algeria and Japan belong in one group, whereas Argentina, Mexico and India belong in another group. Specification 5 and 6 are based on the idea that the effect of the regressors is more important in low income countries. Moreover, in these specifications the standard regressors matter less in the richer countries, but others (either *real exchange rate distortions* (rerd) or *higher education enrollment rates* (h60)) matter more.

Specification 7-10 are based on the more general idea that some regressors may affect growth in nonlinear ways. Specification 7 and 8 use a variety of interaction terms. Specification 9 is based on Kalaitzidakis et al. (2001) who observed that economic growth and education may interact in a very odd, nonlinear way. Namely, for the poorest countries enrollment rates in education have a negative effect, a positive effect for medium rich countries and no effect at all thereafter. Using education variables in three of the four variables is inspired roughly on the literature on the different effect that education may have on growth (e.g. Lucas, 1988; Benhabib and Spiegel, 1994; Kalaitzidakis et al., 2001).

In total, 5 heterogeneous specifications (2-6) and 4 specifications with added nonlinearites (7-10) were used. Every specification exists in a version of model size 3, 7 and 14. So, the five hetegenous equations, and four equations with added nonlinearities, multiplied by the three model sizes, result in a total of 27 different specifications (or datasets); 15 heterogenous ones and 12 with added nonlinearities. Like in experiment 1 and 2, every specification was simulated a number of times; namely, 40 times for each nonlinear equation; and 120 times for each baseline. However, for the BMA method, again for lack of computing power, the datasets examined were reduced by one eighth.

The Hendry and Castle (2012) procedure to deal with nonlinearities was added to the experiment with some slight adjustments. In order to simplify the procedure a little bit, and make it slightly easier for the method to detect nonlinearities, the following equation is used rather than equation (15):

$$y_p = a + \sum_{k=1}^{n} x_k \widehat{\beta_k} + \sum_{k=1}^{n} \sum_{i=1}^{k} x_i x_k \widehat{\beta_k} + \sum_{k=1}^{n} \delta_k 1_{(k=p)} + u_i \quad (19)$$

This means that no higher order functions were included that were present in the DGP[9]. Moreover, in line with the bidding to take a small significance level, the procedure is run with a significance level of .001. Furthermore, the non-linear variables were double-demeaned (as described in Castle and Hendry, 2012).

Applying the algorithms

The main methods examined in the experiments (exclusively in the first two, and partly in the third) are *Gets*, WALS and BMA. The WALS algorithm was written into stata code by Luca and Magnus (2012). The *Gets* algorithm was originally written by Hoover and Perez (1999; 2004), but was created into a Stata module quite recently by Damian Clarke (2013). The BMA algorithm was run in R, which has a module to run this program, which is called BMS (Bayesian model sampling). Zeugner (2012) provides a useful guide. The reason for selecting these algorithms is that BMA and the *Gets* algorithms are the cornerstones of two modeling schools. And while BMA is computationally heavy, WALS and *Gets* are computationally light and are therefore easy to use. BACE (Sala-i-Martin et al., 2004) was not used because it is similar to BMA, and computationally equally heavy. In a second step (experiment 3), the *Autometrics* algorithm was used that is part of Oxmetrics 6.

In BMA, Ley and Steel's prescription is followed for both the g-priors (benchmark priors) and m-priors (random model priors; Fernandez, et al. 2001b; 2009). Ley and Steel (2009) make a recommendation specifically for the purpose of cross-country growth regressions. The MCMC algorithm is run with 2,000,000 iterations and 100,000 burn-ins. While this made it relatively time consuming to run BMA (roughly 10 minutes every time), it reduced the risk of sampling error. It is relatively standard in the literature to use a posterior inclusion probability (henceforth pip) cutoff of somewhere either 10% or 50% as indicative of the good evidence for the relations between this variable and the variable of interest. The 10% cutoff is used, after examination of the results showed that the statistic is relatively conservative. The WALS estimator provides a statistic that is similar to the t-statistic, t. We can call this statistic t* in order to distinguish it from normal t statistics. Unfortunately, t* does not follow the standard normal distribution, but a non-converging Laplace distribution. This makes it difficult to provide p-values, and thereby to select a cutoff point. However, Magnus et al. (2010) and Magnus and Luca (2012) claim that the cutoff of t*=1 corresponds roughly to the equivalent of BMA's pip of 50%, and propose to this as a cutoff value. However, as will be discussed in more detail, this cutoff value appeared much too liberal, and a more pragmatic choice was made at t=1.2. The *Gets* algorithm written for stata that was used followed Hoover and Perez (1999; 2004) closely. While the program allows to adjust the stringency choices, Hoover and Perez are closely followed in selecting the cutoff value of t=1.96 (or p = 5%).

Evaluation

In this paper, Hoover and Perez (2004) are followed in evaluating the results by means of real size and real power ratios. The real size ratio is the achieved size - or as Hendry and Krolzig

[9] This is not regarding the complex relation between education and growth of specification 9, based on of Kalaitzidakis et al. (2001). This relationship cannot be captured by a second order Taylor expansion. However, neither could it be by a third order Taylor expansion.

(2003) say: gauge - divided by the standard cutoff value of significance (.05). In other words, the chance of selecting a variable that does not actually belong in the DGP is divided by the expected percentage of wrong selections in a standard single t-test. A value of 1 corresponds to the a size exactly equal to those of standardized tests, and the lower the value is, the better the method performs.

A crucial issue with the evaluation of ASA's is that their efficacy is that their efficacy may depend on the context. In a dataset in with little noise, much data, few variables, and little multicollinearity, finding the correct set of variables is relatively easy. Any appropriate search algorithm will manage to achieve a low size and a high power, while size is likely to be high and power low in case of noisy data. In order to have comparable standards of evaluation nevertheless, use is made of Hoover and Perez' (2004) real power ratio measure. The real power ratio is the achieved power in the experiment - the number of correctly selected variables – divided by the achieved power of a regression containing all the true variables. A value of 1 corresponds to the power level that is exactly equal to the level of power we achieve when we know the true specification from the start and use statistical methods to test it. In many cases, especially when the signal-to-noise ratio is low, running a test with the true specification from the start, might still result in insignificant coefficients (and a lower than perfect power). Real power measures the achieved power relative to the power we would achieve in the perfect case that we guessed the functional form correctly from the start, and use the statistical test to falsify these ideas. Using this measure ensures that the power measure does not merely capture the signal-to-noise ratio, rather than the efficacy of the method.

Unfortunately, this measure also has some disadvantages. It is a relative measure, and only tells us something about the potency of the method vis-à-vis the potency of standard regression techniques when all the rights regressors are known. Sometimes we might be more interested in absolute measures. For instance, if the signal-to-noise ratio is very low, neither the benchmark regressions nor the search methodologies will provide a higher probability to select the right variable than selecting the wrong variables. At this point, the real power ratio measure will be high, even though the methods do not really work very well. Therefore, the chance of selecting correct variables and the chance of selecting wrong variables is reported too. These measures are referred to as absolute measures, while the real power and size measures are referred to as relative measures.

The evaluation of the nonlinear *Autometrics* procedure is more complex, because due to the large amount of potential variables that can be selected, exceeding the original set of potential regressors by factor 22.5, the results are difficult to compare with the results of the linear ASA's. Firstly, this is simply due to fact that the chance of selecting a correct regressor is much smaller in a larger set of regressors. Secondly, the algorithm might get the regressor correct, while the functional form is misspecified, which is not an option in case of the linear ASA's. Therefore, we have to look at the results a little differently. The relative results are not reported, but merely the absolute ones are.

4. Results

First experiment

The results of the first simulation experiment are summarized in table 2. For easy comparison, Hoover and Perez' (2004) main result table is appended (note that they both contain results for *Gets*). As we can see, and as could be expected, the results show that there is a real power and size tradeoff. While BMA tends to have a very low size, it also is a little less successful identifying the relevant variables. However, while both WALS and *Gets* have an inflated size (larger than .05, indicated by values larger than 1), *Gets* does outperform WALS in that it achieves a very similar real power ratio, but a much lower size. G*ets* does have an increasing size ratio over the model size, which is not the case for WALS. While BMA has a relatively low power in respect to the other two methods, the low power ratio it achieves is compensated by an impressive size. Most notably, its size is exactly zero in case of model size 0. This is in fact quite impressive.

It is interesting to see that the found results on *Gets* are different from Hoover and Perez' results. While the power that *Gets* achieved is roughly similar to the power achieved in Hoover and Perez, the size is generally roughly twice as in our case[10]. This highlights the non-universality of results such as these.

Table 2: Results of experiment 1 (and Hoover and Perez (2004)'s main result table[a])

	WALS		*Gets*		BMA	
model size	real size	real power	real size	real power	real size	real power
0	2.84		1.8		0	
3	2.85	0.73	2.14	0.76	0.61	0.58
7	2.98	0.95	2.22	0.93	0.65	0.7
14	2.87	1.05	2.39	1.05	0.94	0.64

TABLE I

The efficacy of three search algorithms

	Extreme-bounds analysis		*Modified extreme-bounds analysis*		*General-to-specific*	
Models with:	*Size ratio**	*Power ratio†*	*Size ratio**	*Power ratio†*	*Size ratio**	*Power ratio†*
0 true variables	0.060		1.10		0.75	
3 true variables	0.003	0.43	5.17	0.77	0.77	0.95
7 true variables	0.030	0.13	5.89	1.10	0.81	0.93
14 true variables	0.020	0.04	5.45	0.67	1.02	0.82

[a]The results from Hoover and Perez were based on a very similar experiment using 30 datasets per model size (rather than 10 in our case), but each dataset contains the same number of simulated variables (100). Size ratio and power ratio refer to the same concepts as real size and real power in my terminology. In fact, these concepts are borrowed from their paper.

Table 3 shows the absolute statistics. While the information in the table can be reduced from table 2 and the benchmark regression summary (see appendix, A2), it does provide a useful insight on the results that these ASA's really achieve. For each ASA, the first column summarizes the probability that a selected variable is truly part of the DGP given that it is identified as such by an ASA. This is the inverse probability of the achieved power measure (the probability a true variable

[10] Hoover and Perez (1999) explain that the size ratio is likely to be affected by the multicollinearity of the dataset in which the method is used

is selected). For all methods this goes up with model size. This is expected, as in the case of a larger number of variables that matter, it is more likely to select one that matters. However, what is quite impressive of the BMA method in particular, is that this value is quite high even in the small model size case. This means that, while BMA does not quite select all the relevant variables, the ones that it does select are quite likely part of the DGP in our simulation experiment. In many applications this will be quite important. In case of a small model size, *Gets* and WALS are expected to propose models of which 80% of the variables they include are misidentified variables. However, in the case of BMA this value does not become much higher than 50%.

Table 3: Absolute statistics

	WALS(1.2)		gets		BMA	
model size	Prob (true|selection)	Achieved power	Prob (true|selection)	Achieved power	Prob (true|selection)	Achieved power
3	0.179	0.404	0.230	0.416	0.477	0.377
7	0.326	0.361	0.392	0.357	0.511	0.303
14	0.549	0.350	0.597	0.353	0.611	0.284

Robustness to cutoff values

In order to see to what extent the efficacy of the methods is affected by the cutoff levels they employ, a number of different values were tried for all the assessed methods. In case of WALS and BMA, the cutoffs can be set post-hoc. This makes it easy to compare the way these methods react to more stringent, or more liberal cutoff values. Table 4 summarizes the size-power tradeoff for WALS with a model size of seven. At the "natural" level of t*=1, the size is higher than 20% (4.55 * .05), meaning that variables that are not part of the DGP have quite a large probability to be included. At the same time, the real power is not higher than 1. While Magnus et al. (2010) and Luca and Magnus argue that a cutoff point of t*=1 would correspond roughly to the equivalent of BMA's posterior inclusion probability of 50%, the inflated size related to this value seems undesirable. This is how the decision is motivated to use the slightly more conservative cutoff value of t*=1.2 for WALS throughout the paper. Moving from t*=1, to t*=1.2 seems to decrease the size quite a bit without reducing power all that much. The size can be reduced more drastically by accepting a lower power. At the conventional cutoff of 1.96 (which does not have a similar meaning in this case as in case of normal t-tests), real size is much lower than 1, but this comes at the cost of a real power ratio that is also much lower than 1. What becomes clear, is that for the present purpose, WALS is strictly speaking dominated by *Gets* and BMA jointly. If one prefers a high power, *Gets* is more potent than WALS at the higher tradeoff values, while BMA is more potent at the lower size levels.

Table 4: size power tradeoff in WALS (7 variable model size)

t*	Real size	Real power
1	4.555	1.184
1.2	2.975	0.955
1.5	1.454	0.680
1.8	0.678	0.471
1.96	0.431	0.383

While BMA managed to achieve a very good size-power ratio, the power was still relatively low. For that reason, it was assessed whether the power could be increased without increasing size too much if the posterior inclusion probability was shifted from .1 to .05. This implies that the standards of testing are made more liberal; i.e. more regressors will be identified as robustly related to growth; where liberal stands in contrast to stringent: identifying fewer variables as related to growth. The results are summarized in table 5. While the real power of the method is indeed

increased, the costs in terms of size are relatively high in case of the larger models sizes.

Table 5: BMA, employed with a .05 posterior inclusion probability as a cutoff value

model size	real size	real power
0	0	
3	0.64	0.67
7	1.16	0.83
14	1.79	0.86

In case of *Gets*, the cutoff value cannot be altered post-hoc and the simulation of experiment 1 was repeated with a cutoff value of t-values shifted from 1.96 to 2.24 (corresponding to a two-tailed p-value of 2.5%, rather than 5%). The results, summarized in table 6, indicate that the size of *Gets* can be significantly reduced by making the test more stringent. This does come at the cost of power, which is lower than in table 2. Whether this power-size tradeoff is more desirable than the original one is open for discussion. In the remainder of this article, Hoover and Perez (2004) are followed and a cutoff value of 1.96 is used.

Table 6: *Gets*, with a cutoff value of t = 2.24

model size	real size	real power
0	0.94	
3	1.19	.67
7	1.30	.79
14	1.56	.89

Differences with Hoover and Perez (2004)

In general, the results show that BMA is a stringent, but reliable method, and *Gets* is more liberal and powerful. With respect to Hoover and Perez (2004), these results show that the advantage *Gets* has with respect to the Bayesian methods does not translate to the same advantage with respect to BMA. Furthermore, it is interesting to see that quite different results with respect to *Gets* were found than those in Hoover and Perez. Firstly, it may be suspected that this was due, at least partly, to the number of regressors included in the research (34 for Hoover and Perez, 42, for this experiment). For this purpose the experiment was redone from the start with a set of 34 regressors. The results are summarized in table 7. As we can see, in case of *Gets*, real power and size lie somewhat closer together – i.e. the results improved – but the size is by no means close to the level 1 as observed in Hoover and Perez. In fact, the size ratio appears to be quite stable. As we shall see in table 9, the size of *Gets* is quite stable even when moving to datasets with much larger sets of regressors.

Table 7: a repetition of experiment 1, with 34 rather than 42 regressors

	WALS		*Gets*		BMA	
model size	Real size	real power	Real size	real power	Real size	real power
0	2.871		1.894		0.118	
3	2.881	0.835	1.968	0.789	0.761	0.714
7	3.004	1.024	2.060	0.980	0.704	0.689
14	3.192	1.175	2.193	1.076	1.430	0.724

Therefore, the most likely cause is probably that Hoover and Perez (2004) use a very different dataset. For their experiment, the Ferandez et al. (2001b) dataset was used, which contains different variables and is less rich in terms of variety of variables than the Sala-i-Martin et al. (2004) data set that was the main source of the dataset used here. It has been observed that there are very

different aspects to the two different datasets, both of which are often used in the growth empirics literature (e.g. Ley and Steel, 2007). Likewise, the large impact data has in the growth economics context has also been discussed (Ciccone and Jarocinski, 2010). Hoover and Perez (1999) warn that the size of their method may be dependent on the multicollinearity, which may be much higher in this dataset than the one used by them. This result thus shows how much context matters for the evaluation of ASA's.

Number of included variables

A final indication of efficacy of the ASA's examined is the correctness of the sizes of the models the ASA's find. Table 8 summarizes the average number of variables that were identified for each method at different model sizes. In case of *Gets* this simply means the number of variables in the final specification. For WALS and BMA this means how many variables were identified as robust. We can see that all models have a tendency towards medium sized models and therefore underestimate the number of variables. BMA is most flexible in this respect, but also most conservative, and underestimates at all model sizes. WALS is least flexible and has a tendency to overestimate the model size for models of size 0 and 3, and to underestimate the model size for models of size 7 and 14. *Gets* lies in between. It is more flexible than WALS, but still underestimates large models and overestimates small ones.

Table 8: Achieved model sizes

	Gets		WALS		BMA	
model size	mean	st.dev.	mean	st.dev.	mean	st.dev.
0	3.78	2.35	5.96	3.37	0	0
3	5.42	2.56	6.78	3.38	2.37	1.65
7	6.38	2.9	7.73	3.6	4.15	1.98
14	8.28	2.52	8.92	3.38	6.42	2.6

Second experiment

While the results of the first experiment provide reason to be optimistic, the smaller size of the potential set of variables, the easier it is to find the correct variables. In reality, the relative success of the experiment with 42 variables does not provide, by itself, a reason to be optimistic in case of 67 variables. The most common dataset used for this purpose, the SDM data set consists of 67 variables, which is still a small number compared to the 145 variables that are found by Durlauf et al. (2005). It is therefore important to see how the efficacy of the methods responds to an increase in variables to be considered.

The results of the second experiment are summarized in table 9 and 10. While the results in table 9 still look relatively similar to table 2 for the smaller model sizes in WALS and gets – in fact, they look better looking at the real power – looking at the 14 variable size models, the relative potency of the methods go down drastically. The change is even more obvious in the case of BMA. BMA simply does not pick out any variables anymore. While its size is very low, its success to select the correct variables is very small.

In fact, the success of WALS and *Gets* is greatly exaggerated in table 9, as can be seen in table 10. The key is in table A2 in the appendix. The signal-to-noise ratio is apparently much lower in case of the 67 variable case, especially in case of the smaller models, the real power measure may be quite high even if the achieved power is quite low. In table 9, the real power is indeed quite high in *Gets* and WALS for the 3 and 7 model size case. This does not mean that the methods are doing so well, but that the benchmark regressions perform poorly. In the 14 model case, the power is not so low in the benchmark case, and the real power drops dramatically.

Table 10 thus provides a richer picture of what is going on. If we compare the fraction of successful selections to those in table 3, we can see that they have gone down quite drastically, especially for *Gets* and BMA, and for all of them in the 14 variable case. This is in fact quite

remarkable. In case of the 14 variable (14/67 =) 20.9% of the variables belong to the true DGP. For all three specification selection algorithms the probability that a variable is true given that it has been selected is very close to this value too. In other words, a random variable selector might perform very similarly. For BMA and *gets*, the same probabilities for the 3 and 7 variable models are even below (7/67 =) 10.4%, and (3/67 =) 4.4% respectively. In other words, only WALS is able to beat the odds for the small models.

So, how do we interpret this result? There are two possible causes that may contribute to the deterioration in results with respect to experiment 1. Firstly, it may be the case that the due to the fact that data was not so nice – i.e. included missing data. Quality of the data may make a large difference in the ability of statistical method to make solid inferences. What more though, is that most algorithms simply do worse when the amount of variables to search is larger. As variables often are not independent of one another, he true variables will have a harder time to "stand out" from the rest. They "drown" in the pool of variables that partly explain the same information. While we cannot distinguish these two effects clearly, quite likely they both play a substantial role

Table 9: Results of experiment 2 (relative)

	WALS(1.2)		*Gets*		BMA	
model size	real size	real power	real size	real power	real size	real power
0	3.633		1.898		0.000	
3	3.221	3.847	1.854	1.951	0.000	0.000
7	3.140	1.918	1.900	1.360	0.001	0.016
14	3.314	0.672	2.608	0.278	0.006	0.022

Table 10: Results of experiment 2 (absolute)

	WALS		Gets		BMA	
model size	prob (true|selection)	achieved power	prob (true|selection)	achieved power	prob (true|selection)	achieved power
3	0.289	0.171	0.042	0.086	Na[a]	0.000
7	0.472	0.155	0.042	0.039	0.045	0.003
14	0.213	0.173	0.224	0.131	0.182	0.006

[a] No variables were selected, and hence this probability cannot be calculated.

This result is quite important for the application to cross-country growth regressions. A first place, for instance, where this may have gone wrong is a Hoover and Perez' (2004) analysis. While their aim was mostly to argue that the skeptical attitude towards *Gets* on the basis of concerns of data mining was misplaced. However, they do draw a unsupported inference with respect to the application of their method to real world data, and some mild criticism may be in place. They first do a simulation experiment on the efficacy of three specification search algorithms, including *Gets*, that are based on a 34 variable dataset. Then, they assess the method on a 62 variable dataset with messy, and missing, data (using a similar dataset as used here, from Sala-i-Martin, 1997). The fact that *Gets* did very well in the 34 variable case, does not mean at all it is plausible that it does well in the 62 variable case. This inference is simply one that cannot be made, as can be seen in the present case.

Interestingly, while *Gets* seems to do much better than WALS in the 42 variable case, in the 67 variable case, WALS does better as *Gets* deteriorates drastically. The same applies to BMA vis-à-vis WALS. We thus need to be careful in making inferences from success in our smaller, and nicer, data samples to larger and messier ones. The good news though, is that WALS turns out to be a method that deals with all these issues best. While its results also deteriorate in moving from experiment 1 to experiment 2, the problems are less drastic, and the size ratio goes up only marginally. Even though the size ratio of WALS from experiment 1 was larger than those of the rest, the changes in success are clearly a lot smaller, and WALS therefore proves itself quite a robust

method.

The third experiment

Table 11 reports the relative results of the third experiment for the methods used before. As we can see, the results from the baseline regression show slightly different patterns than experiments 1. While the power is higher for all three model sizes, the size is also quite a bit higher. This difference is quite likely due to the fact that the contingencies in the data and only 1 specification was used for each model size (see table 1). In table A3 in the appendix the power of the benchmark regressions is summarized. As we can see, the power is relatively high in the baseline for a model size of 3 and 7, but is very low for a model size of 14. Hence, it is really quite impressive that the real power is still close to unity for model size 3 and 7 for WALS and *Gets*, but less so for a model size of 14. The size is quite a bit higher overall, in particular for BMA of model size 3. It is important though, to consider table 12 as well, as the results in 10 are affected quite substantially by the different results found in table A3.

The important part of table 11 is in the second and third block, where the results of the nonlinear specifications are summarized. The results show the average efficacy of the different nonlinear specifications that are described in detail in table 1. Interestingly, the results of the heterogeneous equations are not that different from the baseline equations. This is particularly the case for WALS and *Gets* for models of size 3 and 7. Real and absolute power, however, do decrease significantly for model size 14 in case of both heterogeneities and added nonlinearities. The overall BMA results do seem to be affected to a substantial extent by nonlinear functional forms. Moving from the baseline to heterogeneous specifications and then to specifications with added nonlinearities increases size, decreases power and the probability that the selected variables are correct. WALS and *Gets* though, appear to be quite robust to nonlinear functional forms, for at least DGP's with a moderate model size.

Table 11: Results of experiment 3 of the main methods (relative)

baseline	WALS		*Gets*		BMA	
model size	real size	real power	real size	real power	real size	real power
3	3.8	1.19	3.08	1.09	1.78	0.75
7	3.2	0.92	2.2	0.84	0.19	0.46
14	3.09	2.64	2.29	2.39	0.57	1.87
heterogeneity[a]						
model size	real size	real power	real size	real power	real size	real power
3	2.87	0.94	2.25	0.92	1.78	0.63
7	3.19	0.93	2.47	0.91	0.64	0.62
14	3.24	1.06	2.57	1.05	0.86	0.71
Added nonlinearities[a]						
model size	real size	real power	real size	real power	real size	real power
3	3.4	1.22	2.84	1.09	1.72	0.42
7	3.32	0.92	2.74	0.82	1.2	0.32
14	3.07	1.58	2.15	1.42	1.25	0.55

[a] The measures on the heterogeneous equations and nonlinear equations in this table represent averages of the different specifications described in more detail in table 1.

Table 12 Results of experiment 3 of the main methods (absolute)

baseline	WALS		*Gets*		BMA	
model size	pr(true| included)	achieved power	pr(true| included)	achieved power	pr(true| included)	achieved power
3	0.19	0.56	0.21	0.52	0.24	0.24
7	0.37	0.48	0.44	0.44	0.83	0.23
14	0.53	0.35	0.58	0.32	0.81	0.04
heterogeneities[a]						
model size	pr(true| included)	achieved power	pr(true| included)	achieved power	pr(true| included)	achieved power
3	0.2	0.44	0.24	0.44	0.21	0.21
7	0.38	0.48	0.44	0.47	0.65	0.24
14	0.59	0.14	0.64	0.14	0.76	0.02
Added nonlinearities[a]						
model size	pr(true| included)	achieved power	pr(true| included)	achieved power	pr(true| included)	achieved power
3	0.16	0.58	0.17	0.52	0.12	0.24
7	0.39	0.48	0.42	0.42	0.4	0.22
14	0.54	0.21	0.54	0.19	0.56	0.02

[a] The measures on the heterogeneous equations and nonlinear equations in this table represent averages of the different specifications described in more detail in table 1.

A main question interest is if the methods intended for situations as these in fact do better than standard techniques. The results for Castle and Hendry's method are summarize in table 13. The first block shows the results if the same evaluation standards are used as before: getting the variable exactly right. As we can see, the results are adverse in all directions. Even though the size is relatively low, the achieved size and percentage of selected variables that are correct is low. For the heterogeneous specifications and nonlinear specifications with 7 variables, the power is slightly higher, but the size is high too. In other words, the method, as seen in this light, selects many variables that are strictly speaking incorrect and few that are strictly speaking correct.

However, even though a double de-meaning technique was used in order to avoid strong collinearities between levels and their alternative functional forms, it was found that in many regression the true linear regressors were included in a nonlinear way. Therefore, the success of the method to identify cases in which the regressor was correct, but wrongly specified, are reported too. The results of this evaluation are summarized in the second block of table 13.

Table 13: Results for Hendry and Castle's (2012) method

perfect									
	baseline			heterogeneity[a]			added nonlinearities[a]		
model size	prb(truth|selection)	achieved power	Real size	prb(truth|selection)	achieved power	Real size	prb(truth|selection)	achieved power	Real size
3	0	0	0.03	0.02	0.01	0.05	0.01	0.02	0.1
7	0.07	0.09	0.07	0.08	0.13	0.1	0.04	0.1	0.14
14	0	0	0.09	0.06	0.03	0.12	0.05	0.04	0.21

correct, but wrong functional form									
	baseline			heterogeneity[a]			added nonlinearities[a]		
model size	prb(truth|selection)	achieved power	Real size	prb(truth|selection)	achieved power	Real size	prb(truth|selection)	achieved power	Real size
3	0.89	0.38	0.07	0.47	0.26	0.62	0.28	0.27	1.74
7	0.64	0.34	0.76	0.63	0.39	1.07	0.57	0.46	1.8
14	0.79	0.25	0.67	0.82	0.34	0.8	0.71	0.44	2.18

[a] The measures on the heterogeneous equations and nonlinear equations in this table represent averages of the different specifications described in more detail in table 1.

What is striking about this method is that the power, probability of true variable selection, and real size, are all very promising. In the baseline 3 variable model case, almost 90% of all the included variables were at least in some way related to the true regressors in the DGP. What is unfortunate though, is that the impressive results of the baseline do decrease in the specifications with added nonlinearities. While for the heterogeneous specifications the results are quite impressive, for specifications with the added nonlinearities the probability that the selected variables are correct is lower and size much higher.

While some of these results seem promising, optimism has to remain moderate in light of the fact that functional form incorrectness may be a serious problem for inference. If variable linearly affects growth, but is taken by the method to affect growth exponentially or in an interaction with another variable, this knowledge may be quite misleading in policy applications. On two occasions, this ASA got the nonlinearity exactly right, which is an achievement that linear ASA's could never share. At the same time, this number is quite low. Still, the total set of relevant regressors is 129 for the 3 variable model size case, 278 for the 7 variable case, and 525 for the 14 variable case. This means that 13.7%, 30.3% and 55.6% of the variables would expected to be identified correctly if variables were selected at random. In all cases, except the 14 variable added nonlinearities case, the success percentage of the methods lies much higher than this.

In general, the conclusion has to be that the method is quite successful in selecting relevant variables, but not so much in selecting the correct functional form. Moreover, we can conclude that the method still has severe issues in case of specifications where non-linear terms play an important role. Given that this method was specifically designed for the purpose of nonlinear model selection, this may be considered a serious problem.

Overall results

As the results from experiments 1 indicate, specification search algorithms can be quite potent. While BMA is quite good at selecting a small subset of variables of which many are truly related to growth, it often leaves quite some relevant factors out. Both *Gets* and WALS are relatively good at selecting a larger set of variables which include a large share of the relevant variables (though *Gets* does this a little better than WALS). In fact, the real power that these methods are able to achieve lies close to the bench line level: the level one would achieve if the true specification was already known a priori (!). However, this does come at the price of an inflated size, especially for WALS.

The inflated size of *Gets* found is higher than the size Hoover and Perez find for *Gets* in their experiment on a different dataset. This illustrates the importance of context. Together with the disappointing results in experiment 2 – with the full 67 variable set - this shows that the efficacy of ASA's is highly dependent on the quality of the data and number of variables to be researched. WALS was most robust to these changes, while the change in number and quality of the regressors in experiment 2 proved a particularly pertinent concern for BMA. BMA hardly selected any variables in the 67 variables case, but still fails to select the correct ones. In case of added nonlinearities, or parameter heterogeneity, the results did not change so much with respect to the baseline case. However, for the larger variable case, this result does not hold anymore. BMA turns out to be least robust to nonlinear functional forms. Lastly, the *Autometrics* algorithm intended for nonlinear specification searches performed quite poorly in selecting the correct specifications, but turned out to do quite well in terms of selecting the right variables with the wrong specification.

In summary, we can conclude that *Gets* is a successful liberal search strategy, BMA is a successful stringent search strategy, and while WALS is less successful than the other ASA's in experiment 1, it is most robust to changes in functional form and data quality.

5. Real growth data

So far, only simulated growth regressions have been evaluated. In this section the knowledge gained in the former section is applied to real growth data. The results of the ASA's applied to real growth data are summarized in table 14. Because all methods did so much worse in the 67 variable case, the reduced dataset of 42 variables was used. In case of the nonlinear *Autometrics* methods, not all results are reported, but it was simply reported how many variables were mentioned at least twice in the final specification in one of the nonlinear terms (the specification the method arrived at only included nonlinear terms). This is due to the fact that it turned out to be quite unsuccessful in functional form selection (see experiment 3). The full results of the nonlinear *Autometrics* algorithm applied to real growth data are reported in appendix table A4. For BAT, BMA and WALS only the results are reported that exceed the used cutoff value (t*=1.2 for WALS, pip=.1 for BAT and BMA).

While half of the variables in the dataset are mentioned at least by one method (21 out of 42), there are a number of variables that come out most often. In assessing the results, we must keep in mind that BMA is a much more stringent method than the other four, and while it reports only 3 variables to be robustly related to growth, the chance that these variables are falsely identified as robust growth determinants is smaller than in case of the other methods (for BAT we do not have good evidence for this; and for nonlinear *Autometrics* this is complicated to assess).

Table 14 shows that, there is no inconsistency in terms of signs (that is, one ASA identifying a variable as robustly positively related, while another showing evidence for a positive relationship). In general, there appears to be good evidence for what is called the conditional convergence hypothesis in the data, considering our methods (see Sala-i-Martin, 1996): three out of our five methods imply a negative coefficient on *initial GDP per capita* (gdpch60l). The same applies for *real exchange rate distortions* (rerd). Moreover, given the stringency of the BMA method, we can say that there is strong evidence for the number of years *an economy has been open* (yrsopen; which is also mentioned another time besides by BMA), *and fertility rates in the 1960's* (fertldc1) as well. Furthermore, moderate evidence is found for the covariance of *investment price* (iprice1), *share of GDP in mining* (mining), *fraction speaking other language* (othfrac), share of *Buddhist* and *Confucian population* (Buddha and Confuc) with long-run economic growth.

These findings are first of all compatible with the view that convergence occurs, barriers to trade and investment matter (years open economy, real exchange rate distortions, investment price), natural resources matter (fertility rates and mining), and religion matters (Buddha and Confucian)[11]. The fraction of people that speak a foreign language may be compatible with a number of growth hypotheses, like stressing the importance of education, globalization or reflecting a history of trade.

Because all methods are not very accurate if it comes to model size, it is hard to determine how many regressors a true model of economic growth would exist of. However, this is the kind of result that could have been brought about by a 7 variable sized model, if the DGP is similar to the ones that were simulated in experiment 1 (table 8). In this case, we can expect roughly 33% of the variables from WALS, 39% of the variables from *Gets* and 51% of the variables from BMA to belong to the true DGP if it is mostly linear (see table 3). While this is still a low number perhaps for WALS and *Gets*, it does show that we can conjecture that a good share of the doubly identified variables are indeed determinants of growth[12].

[11] However, a number of people (like Angrist and Pischke, 2010, for instance) seem to take a skeptical view towards findings like this. It seems to make more sense to interpret these coefficients as representing the presence of the up and coming Asian tigers in the past 50 years. Both interpretations are compatible with the evidence, and the data does not itself lend itself better to any one of these interpretations.

[12] We have to keep in mind though that while the methods are different they do use the same information. It is therefore not fully correct argue that a variable that is identified in gets and WALS, has a probability of 1-(1-.33)(1-.39) = .59 to be correct.

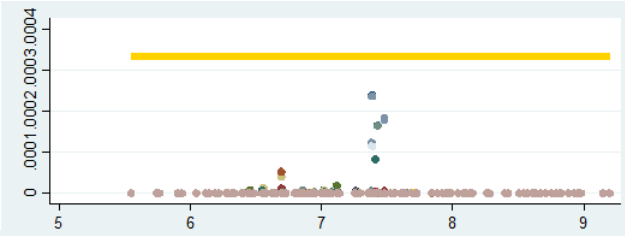
BAT

While the BAT methodology has not been assessed by means of Monte Carlo experiments, the way how it deals with nonlinearities is quite useful in light of the worries voiced about the linear restrictions of the other methods. In order to do so, the 42 variable dataset was used, and four variables are identified as containing potential threshold nonlinearities. The choice of these variables was simply made on the basis of theory, and *were initial GDP per capita*, an *openness* measure (which are both also used in Cuaresma and Doppelhofer, 2007), *political rights* (which is key in new institutional economics; e.g. Acemoglu et al., 2001), and *primary school enrolment rates* (for education; e.g. Kalaitzidakis et al., 2001). In figures 1, 2, 3 and 4 we can see the results of the estimated nonlinearities, and the corresponding estimated thresholds probabilities. As can be seen, not a single observation reaches the prior inclusion probability threshold. While some evidence is found for nonlinearities in initial GDP per capita between log values of 7 and 8, as was also observed by Cuaresma and Doppelhofer, the other nonlinearities stay far from the threshold. Different though, is that in our case, no nonlinearities for the openness variable are observed. We have to keep in mind though, that the conclusion of whether or not there are threshold nonlinearities is highly dependent on the prior probability, which in turn depends on the number of expected variables that contain threshold linearities (which, like in Cuaresma and Doppelhofer, we set to 1).

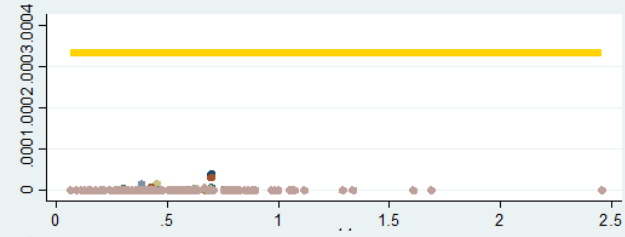
Figures 1-4: nonlinear threshold effect probabilities[a]

Nonlinear variables

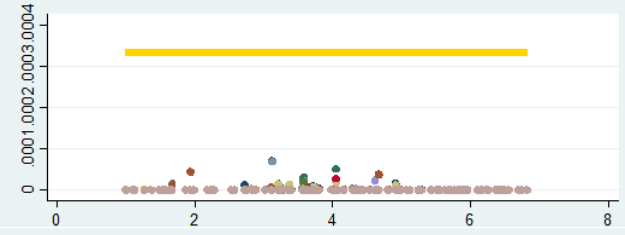
Initial GDP per capita



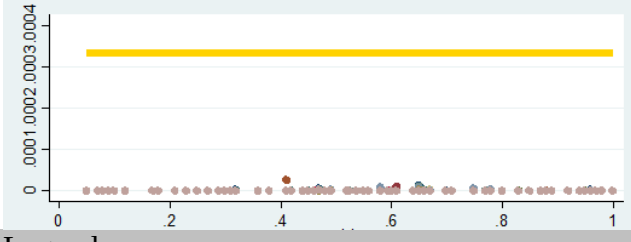
openness



political rights



primary school enrollment



Legend:
x: estimated probability of interaction
y: nonlinear variable
The different colored dots represent different variable interactions with the nonlinear variable. The horizontal line is the threshold level.

[a]The figures show the estimated values of the nonlinear effects. The yellow thick line in all the figures represents the

threshold line of the prior inclusion probability.

Nonlinear autometrics

The full results for the nonlinear version of the *autometrics* algorithm can be found in appendix table A4. We can see that a large number of interaction terms are recommended. We know from the experiment though, that this method tends to overestimate the number of nonlinear relationships (table 13). At the same time, we do know that most of the variables that are identified by this method are expected to contain some degree of truth, in that at least one of the terms from the interaction was related to growth in an alternative functional form. Due to the finding that the nonlinear *autometrics* method resulted in an exaggeration of the number of nonlinear effects only the variables were included in table 14 that appear twice in the full estimation in a nonlinear functional form (no variables were found to relate to growth in a linear way; table A4).

Table 14: results of the methods on the true growth variable

	variables	WALS			BAT			*Gets*				BMA			nonlinear *Autometrics*	total
	gr6010	Coef.	Std. Err.	t*	pip	post. mean	Post. Std. Err.	Coef.	Std. Err.	t	P-value	pip	post. mean	Post. Std. Err.		
1	**gdpch60l**	-.0059	.0029	-2.04	.41	-.00643	.00212	-.008	.0021	-3.8	.000					3
2	**rerd**	-.00007	.000038	-1.85	.57	-.0001	.00003	-.0000697	.000028	-2.47	.015					3
3	**confuc**	.0333	.01834	1.82				.0539	.0136	3.96	.000					2
4	**govnom1**	-.0426	.0261	-1.63												1
5	**brit**	.0056	.0035	1.59												1
6	**othfrac**	.0057	.0037	1.52	.95	.06254	.01528									2
7	**mining**	.0245	.0176	1.39											XXXX	2
8	**iprice1**	-.000031	.000023	-1.36				-.0000489	.0000195	-2.51	.014					2
9	**p60**	.0097	.00719	1.34												1
10	**dens60**	8.98e-06	7.00e-06	1.28												1
11	**scout**	-.0029	.00233	-1.23												1
12	**buddha**				.52	.02704	.00721	.0168	.0066	2.53	.013					2
13	**yrsopen**				.47	.01081	.00398					.14	.00163	.00429		2
14	**life060**				.40	.00042	.00016									1
15	**malfal66**				.37	-.00995	.00322								XXXX	2
16	**fertldc1**							-.0154	.004	-3.89	0	.89	-.0132	.00538		2
17	**muslim00**							.0112	.0034	3.28	.001					1
18	**p60**							.0172	.005	3.4	.001					1
19	**East**											.98	.0224	.00495		1
20	**troppop**														XXXX	1
21	**brit**														XXXX	1
	model size		11			7			8				3			

[a]The variable name references can be found in table A1 in the appendix. In case of WALS and *Gets* the coefficient is the estimated effect of the variable, and the Std.Err the related standard error of estimation. t* refers to Magnus et al.'s (2010) pseudo-t-statistic, a cutoff value of 1.2 is used. For Gets the t value and related p-value are reported. For BAT and BMA the posterior inclusion probability of each of the variables that reach the .1 cutoff value is reported, along with the relevant posterior mean and posterior standard error. In case of the nonlinear *Autometrics* estimation, the XXXX's refer to variables that are mentioned at least twice in a nonlinear fashion in the full estimation results (reported in appendix: A4).

6. Discussion:
Consequences for growth, and consequences for methodology

Consequences for growth methodology: contextualism

The main conclusion of the conducted experiments is that context matters. We cannot unambiguously say that one specification method works, whereas another does not. The efficacy of the methods relies very heavily on the number of variables included, the (expected) model size, the signal-to-noise ratio of the variables, and on the functional form of the DGP. Whereas the achieved results of the experiment with 42 variables was not as promising as the best results of Hoover and Perez's (2004) experiment with 34 variables, the results for the 67 variable case were quite adverse. Hence, on the basis of the limited robustness of ASA's to the number of potential variables, a recommendation can be made:

Recommendation 1: In case little theory is available for variable selection, specification methods are very useful unless too many potential regressors exist. That is, never as large as 67, but the smaller the set of potential regressors, the more reliable the methods are.

This recommendation implies that ASA's can work together very nicely with theory, but, it should be noted that theory should start to be more restrictive rather than constructive: it should tell us which variables are unlikely to matter. The open-endedness of growth research (Brock and Durlauf, 2001; Durlauf et al., 2005) deteriorates the analysis if it goes as far as leaving us with over 50 variables.

However, in case one needs to look at larger sets of variables, WALS is recommended as it achieved the best robustness to size of the set of potential regressor. Still, a warning is at place: the large size of the method does imply that many false inferences are expected to be made. In the end WALS is likely to select many variables that do not belong to the true DGP, and is thus very liberal.

Recommendation 2: In case one wants to look at larger sets of potential regressors, WALS is recommended.

A further important point is that depending on the aim of the research, different methods may be appropriate. As the experiments show, BMA is a relatively reliable, stringent method. While its power was lower than for the other methods, so was its size, and the achieved success rate of the selected variables was highest of all. In case one wants to reduce a set of potential variables to a small set that is not exhaustive, but has a high likelihood of being related to growth, the experiments provide evidence that BMA is the method to be recommended. The experiments provide evidence that *Gets* achieves a higher power, and a more reliable estimate of the true model size, but comes at the cost of a higher size.

Recommendation 3: In case a conservative selection is required, BMA is recommended, in case a liberal selection is required, *Gets* is recommended.

Finally, if it comes to (suspected) nonlinearities, recommendations are hard to make. In the experiments, it was found that in case of nonlinear relationships, *Gets* and WALS still are quite good at identifying the (heterogeneous) linear relationships. While added nonlinearities do distort the success of the methods, the performance of the analyzed methods do not go down dramatically. Unfortunately, we could not lay the BAT approach to experimental scrutiny, but we could do this for Hendry and Castle's (2012) approach. What was found is that while it performed well in terms of identifying the correct regressors, it performed poorly in terms of identifying the correct functional form. Moreover, it seems to have a strong tendency to identify too many nonlinear relationships. In this sense, it is thus not an improvement in terms of functional form identification. We can thus recommend the following:

Recommendation 4: if mild nonlinear relationship are expected to exist alongside linear relationships, and one is mainly interested in the linear components, it is recommended to use the linear methods.

Recommendation 5: If one is particularly interested in nonlinear components, BAT appears to be a rather conservative method. Hendry and Castle's (2012) method, in this context, is very (perhaps overly) liberal. Because a conservative aim might be advisable, using BAT is recommend (but more research is needed).

Consequences for econometric methodology at large (with a historical background note)

Two debates among early econometricians let to a very skeptical account about data-driven econometrics. Firstly, when Jan Tinbergen presented the first statistical model of the United States economy (1939), it started a fierce debate between a number of econometricians about the usage of probability theory. A young Milton Friedman (1940) participated in the debate, and argued that Tinbergen made the crucial mistake to select his determinants of industrial production on the basis of statistical methods, while, at the same time, he used statistical methods to evaluate the model and the plausibility of its determinants. Most prominent though, was Keynes (1939; 1940) debate with Tinbergen (1939; see Louca, 2007; and Hendry and Morgan, 1995). Keynes critique was devastating. Like Friedman, he worried about the reliability of regressors that were chosen by trial and error, the time period that seemed to have been chosen in a data-driven way, and the meaning of regression coefficients in the presence of omitted variables - which are always there, he argued, because many important variables cannot be measured.

A second important influence was Koopmans's (1947) paper on econometrics, which ended up having a large impact on the methodology of the influential Cowles commission. Koopmans argued against theory-free variable selection on the grounds of three arguments: 1) theory is needed for identification of potential regressors, otherwise no inference is at all possible ; 2) theory must be used to identify functional form and interpretation of coefficients; and 3) without theory, any causal interpretation is ambiguous.

All these arguments carry much weight, and for good reasons econometricians were skeptical of data-driven methods for a long time (e.g. Backhouse and Morgan, 2000). An important contribution to this skeptical point of view, for instance, was Lovell (1983), who showed that the analytical result that multiple hypothesis testing leads to an increased size turned out to be quite real. As Backhouse and Morgan claim, this current in econometric methodology was fed by a popperian spirit - that maintained that only falsification leads to solid inferences - and data-based discovery does not fit very well in this framework.

By all standards, it seems that simulation experiments like Hoover and Perez (2004), but also Hendry and Krolzig (2003), and Magnus et al., (2010), show that much of this skepticism is exaggerated. While, surely, some theory is required for good inference, these experiments show that the amount of theory that is required for reliable inference may be quite limited. While repeated testing and data-driven inference may lead to unreliable inferences in many circumstances, if done carefully and in a sophisticated manner, data-driven inferences may be quite reliable in certain circumstances. In fact, in cases where theory is not very helpful, this appears to be the most expedient method of inference (Van der Deijl, 2013).

Two issues remain unresolved. First of all, the conclusion that data-driven inference may lead to reliable results does not yet teach us under what circumstances we can use it or not. To this latter problem, I have intended to make a contribution. We would not only like to know whether data-driven inference may work, but we want to know under which circumstances it does work. This has been discussed above, and I think that it is indeed important to realize that context matters a lot for the efficacy of the methods.

A second issue is one that is not yet resolved, and one for which it is much harder to design experiments like the ones presented above: the causality problem. Whereas many of Koopmans',

Keynes' and Friedman's worries with respect to statistical inference seem to be dealt with quite well in the discussed ASA methodologies, determining causality remains a complex problem. What the experiments show is that if a DGP exists, we can discover a good share of the robust, or true, correlational structure, simply by applying the ASA's. However, the interpretations we should give to these correlational structures is not quite straightforward. We might want to take our findings as evidence that, for instance, speaking foreign languages causes growth, but, this is not in itself information that is contained in the data. The data may be generated by a different relationship, for instance: richer countries invest more in language education. Or, even worse, speaking a foreign language is not in itself the relevant factor, but another factor is, that is related, like being a part of an area in which much trade occurs. This worry is not only shared by these early econometricians, but voiced by a number of more contemporary commentators too (most prominently in Angrist and Pischke, 2010; and see Durlauf and Quah, 1999, for this case in growth economics). This issue touches to core of methodological issues concerning specification searching, and stretches in relevance to many other econometric practices too. What do these econometrical models really tell us about causality, even if the statistical relationships they establish are genuine.

While it is true of the area of research that causality is not implied in the structure it seeks to identify, a few reassuring words may be due.

First of all, knowledge of correlations is useful in itself, in that it can inspire research towards the causal relationships that play a role. Secondly, while it is popular to assert that "correlation does not imply causality", this statement is not quite plausible (Reiss, 2005; 2008). If a correlation is genuine, in the sense that it is not due to mere chance, it implies causality in the sense that if A is correlated to B, either A causes B, B causes A, or there is common cause that causes them both[13] (from Reichenbach, 1956, as quoted by Reiss, 2008). Correlation thus implies that at least some causation is there. This may help to identify causal structures (see for instance Spirtes et al., 2000, for an account of how we can identify causal structures from correlational structures).

Lastly, a serious issue for identifying causal structures is that omitted variables may be the true causes and are left out of the equation (one of Keynes' main objections to Tinbergen, 1939). Hoover and Perez (2004) argue that this is not something to worry about too much in this context. If someone argues that the true cause of a certain correlation is left out, one can simply add it and run it again. This does not, of course, resolve the issue. After all, many important variables may simply be immeasurable. However, if the right context and interpretation is understood, classical worries related to data-driven econometrics may often disappear. In relation to this particular issue: some of the results discussed here have shown that ASA's perform much better than expected under the presence of omitted (nonlinear) variables.

In light of the open-endedness of growth theory research, empirical methods such as the ones discussed here are essential for learning about the data. While much work and knowledge is still to be gained in the path to understanding long-run growth, cross-country growth regressions and ASA's play an important role in this research project.

Future research

As mentioned in the introduction, the two philosophies of inference discussed (General-to-specific and Bayesian inference) have developed into two separate literatures. Given their common interest, this - in the least - odd. In the experiments strengths and weaknesses of both approaches have come to light, and it seems that synergy is to be gained from cooperation. The following discussion is intended to briefly discuss what are taken to be strengths and weaknesses of the different methods, and present a possible way in which we may be able to achieve best of both worlds.

While *Gets* seems to be based on a sophisticated idea of what a true DGP is expected to look like, its output is only based on a single regression specification. It is quite likely that the path that led to these specifications is determined by contingencies in the data that guided the algorithm

[13] Or, needless to say, a combination of these three options occur simultaneously.

towards this single specification. We saw that BMA took a lot of computing power, but ended up being quite reliable in terms of the variables it determined as robust (even though it did not manage to identify a large share of the determinants). The averaging in BMA seemed to be a contributing factor to reliable results. Intuitively, this is because it simply uses a lot more information than can be contained in a single specification. Averaging makes sure that the final result is representative of the information contained in the data. In my view, this is exactly what *Gets* is missing. It may happen in *Gets* that out of 10 specifications that are congruent simplification of a GUM, 9 contain variable A, but, because specification 10 has a slightly higher BIC, variable A is not represented in the output. It seems that an averaging of the models, in this case, will be more informative than model selection based on BIC.

Similarly, it seems to me that BMA could benefit from diagnostic tests. After all, the usage of diagnostics test was intended to determine whether the statistical inferences are valid. Many of the models that are averaged in BMA, and partly determine the outcome, will not meet many requirements that we would want valid statistical inference to meet. Nevertheless, they are weighted in a similar fashion as models that do meet these requirements. Moreover, in light of the partial success of *Gets*, this has proven to do its job quite well.

It seems that combining useful features of the different approaches would seem like a good opportunity to improve the efficacy of the search methods. What is quite interesting is that the aspects of *Gets* that seems to drive its success - taking notice of congruence and diagnostics tests - seem to be perfectly compatible with the methodological aspects that drive the success of BMA, which is the fact that it efficiently makes use of the information contained in a large set of all the possible models that might contain information about the relationship between variables. In future work, possible ways to combine these mechanisms in a unified search algorithm could be assessed. It seems important to me to combine all the aspects of ASA's that could contribute to successful inference, as saturating all the information from the data is what is terribly crucial in data-driven research when data is scarce.

As a starting point, the *Gets* already makes use of an information criterion in selecting a final specification. This mechanism can easily be expanded to include some of the good features of BMA. Rather than searching 1 specification, BMA could be used to assess the likelihood that certain variables are true determinants of all the models that pass all the *Gets* congruence tests. A second possibility is to introduce diagnostic test in the BMA framework, such that models that do not pass (some of) the congruence tests would get penalized. Research in this direction might be able to identify ways to combine the best of both worlds in a single algorithm.

Combining the different strengths of the approaches in one algorithm might be very useful, not only in the growth economics case, but in all cases where data-driven methods need to be applied and data is scarce.

7. Summary and conclusion

For two decades, researchers have been struggling with the question what we can learn about growth from the cross-country growth regressions. As growth theories left much undetermined, applying standard econometric methodology became problematic, due to a lack of theoretical guidance. Researchers turned to econometric methods that did not require much theoretical guidance. This practice evoked much skepticism. A number of researchers have attempted to convince the public of the use of these methods by means of simulation studies (e.g. Hendry and Krolzig, 2003; Hoover and Perez, 2004). In this essay the most up-to-date versions of the econometric methods have been assessed in order to analyze how well they are able to identify true correlational structures in the cross-country growth data, and expanded Hoover and Perez' testing methodology to identify how robust ASA's are to changes in context. While shifting away from the question whether these methods are useful at all, I have tried to find a more specific answer to the question, how useful they are and under what conditions they can be expected to do well. It was argued that in order to truly learn what these methods can teach us, we need to go beyond the kind of tests that were based on linear models with relatively few potential regressors. In the experiments conducted the robustness of the method was assessed with respect to size of the set of potential regressors, data quality, and nonlinearities.

We can conclude that for the application of cross-country growth regressors it is important to keep in mind that size of the set of potential regressors and quality of the data may have major effects on reliability of inference. The presence of mild nonlinearities does not affect the research too much in case research is aimed at the linear part of the structure. The robustness of the efficacy of ASA's to context is highly dependent on which specific ASA is under scrutiny. In general, the *Gets* algorithm was identified as liberal, but powerful, BMA as stringent, but reliable, and WALS as less successful, but more robust than the other two methods.

The variables for which the strongest evidence is found for being importantly related to growth *are initial GDP per capita, real exchange rate distortions, years the economy has been open,* and *initial fertility rates.* The (seemingly conservative) BAT method did not identify any evidence for nonlinearities, while the over-liberal *autometrics* method did identify a large number of nonlinear relationships for *share of GDP in mining, malaria prevalence, share of population living in the tropics and being a former British colony.*

Appendix

Table A1: A summary of the data used (mostly from DSM, 2004

code in tables	full variable name	mean	Included in 42 variable set	Included in 32 variable set
(zz) gr6010	Average annual growth from 1960-2010	0.0170	x	x
openness	Openness measure	74.8647	0	0
zzavelf	Ethnolinguistic fractionalization	0.4786	1	0
abslatit	Absolute latitude	23.1432	1	1
airdist	Air distance to important global capitals	4401.2630	1	1
brit	Former british colony	0.3516	1	0
buddha	Fraction Buddhist	0.0415	1	1
cath00	Fraction Catholic	0.2938	1	1
zzciv	Civil Liberties	0.4777	1	0
colony	Colony Dummy	0.7500	1	0
confuc	Fraction Confucian	0.0133	1	1
dens60	Population Density	151.4180	1	1
dens65c	Population Coastal Density	125.7390	1	1
dens65i	Interior Density	43.5611	0	0
zzdpop6090	Population Growth Rate 1960-90	0.0229	0	0
east	East Asian Dummy	0.0938	1	1
zzecorg	Capitalism measure	3.2857	0	0
engfrac	English Speaking Population	0.0749	0	0
europe	European Dummy	0.1719	1	0
fertldc1	Fertility Rates in 1960s	1.5787	1	0
gde1	Defense Spending Share	0.0253	0	0
gdpch60l	Initial Income (Log GDP in 1960)	7.3166	1	1
geerec1	Public Educ. Spend. /GDP in 1960s	0.0240	0	0
ggcfd3	Public Investment Share	0.0540	1	1
govnom1	Gov C Share deflated with GDP prices	0.1455	1	1
govsh61	Government Share of GDP	0.1639	1	1
gvr61	Government Consumption Share	0.1176	1	1
h60	Higher Education in 1960	0.0311	1	1
herf00	Religion Measure	0.7869	0	0
hindu00	Fraction Hindus	0.0273	1	1
iprice1	Investment Price	93.6353	1	1
laam	Latin American Dummy	0.1797	1	1
landarea	Land Area	803797.5000	0	0

landlock	Landlocked Country Dummy	0.1797	0	0
lhcpc	Hydrocarbon Deposits in 1993	0.7731	0	0
life060	Life Expectancy	52.6909	1	1
lt100cr	Fraction Land Area Near Navig. Water	0.4527	0	0
malfal66	Malaria Prevalence	0.3790	1	1
mining	Fraction GDP in Mining	0.0579	1	1
muslim00	Fraction Muslim	0.2164	1	1
newstate	Timing of Independence	1.1953	0	0
oil	Oil Producing Country Dummy	0.0938	0	0
zzopen	(Imports + Exports)/GDP	0.5538	1	1
orth00	Fraction Othodox	0.0235	0	0
othfrac	Fraction Speaking Foreign Language	0.3016	1	1
p60	Primary Schooling Enrollment	0.6839	1	1
zzpi6090	Average Inflation 1960-90	12.8796	0	0
zzsqpi6090	Squared Average Inflation 1960-90	356.2189	1	1
zzprights	Political rights measure (freedom house)	4.1917	0	0
pop1560	Fraction Population Less than 15	0.3945	0	0
pop60	Population in 1960	21472.8600	1	1
pop6560	Fraction Population Over 65	0.0457	1	1
priexp70	Primary Exports in 1970	0.7438	1	1
prot00	Real Exchange Rate Distortions	0.1287	1	1
rerd	Fraction Protestants	125.8142	1	0
revcoup	Revolutions and Coups	0.2120	1	1
safrica	Sub-Saharan Africa Dummy	0.3438	1	0
scout	Outward Orientation	0.3534	0	0
size60	Size of Economy	15.8934	0	0
socialist	Socialist Dummy	0.1240	1	1
spain	Spanish Colony Dummy	0.1250	0	0
tot1dec1	Terms of Trade Growth in 1960s	0.0059	0	0
totind	Terms of Trade Ranking	0.2580	0	0
tropicar	Fraction of Tropical Area	0.5703	1	1
troppop	Fraction Population In Tropics	0.3107	1	1
wartime	Fraction Spent in War 1960-90	0.0852	0	0
wartorn	War Particpation 1960-90	0.3984	0	0
yrsopen	Years Open 1950-94	0.3354	1	1
ztropics	Tropical Climate zone	0.19	0	0

[a]Variables that start with zz indicate that the variable has been "refreshed" by me.

Table A2: benchmark power		
	42 variables	67 variables
3	0.555	0.048
7	0.389	0.097
14	0.337	0.258

Table A3 benchmark power for the nonlinear specifications			
	baseline	heterogeneity	nonlinearities
3	0.47	0.49	0.35
7	0.52	0.5	0.47
14	0.13	0.43	0.27

Table A4: full results of running Hendry and Castle (2012) method on the true growth variable				
	coef.	st.error	t	prob.
malfal66sqt1	-0.0219	0.0021	-10.50	0.0000
miningXbrit	-0.0641	0.0139	-4.63	0.0000
miningXgovsh61	0.9704	0.2186	4.44	0.0000
miningXgvr61	-1.0755	0.2489	-4.32	0.0000
miningXmalfal66	0.0833	0.0178	4.68	0.0000
muslim00Xmining	-0.1217	0.0177	-6.89	0.0000
othfracXmalfal66	0.0300	0.0058	5.18	0.0000
pop1560Xlaam	-0.0213	0.0044	-4.84	0.0000
rerdXconfuc	0.0003	0.0001	3.51	0.0006
troppopXbrit	0.0160	0.0034	4.73	0.0000
troppopXscout	-0.0141	0.0031	-4.60	0.0000
yrsopenXeast	0.0193	0.0037	5.27	0.0000
r-squared	0.7660	Adjusted r-squared	0.7440	

References

Abramovitz, M. (1986). Catching up, forging ahead, and falling behind. *Journal of Economic history*, *46*(2), 385-406.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *The American Economic Review*, *91*(5), 1369-1401.

Angrist, J. D., & Pischke, J. S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *The Journal of Economic Perspectives*, *24*(2), 3-30.

Baştürk, N., Paap, R., & van Dijk, D. (2012). Structural differences in economic growth: an endogenous clustering approach. *Applied Economics*, *44*(1), 119-134.

Baumol, W. J. (1986). Productivity growth, convergence, and welfare: what the long-run data show. *The American Economic Review*, 1072-1085.

Benhabib, J., & Spiegel, M. M. (1994). The role of human capital in economic development evidence from aggregate cross-country data. *Journal of Monetary economics*, *34*(2), 143-| 173.

Brock, W. A., & Durlauf, S. N. (2001). What have we learned from a decade of empirical research on growth? Growth Empirics and Reality. *World Bank Economic Review*, *15*(2), 229-271.

Campos, J., Ericsson, N. R., & Hendry, D. F. (2005). General-to-specific modeling: an overview and selected bibliography.

Cass, D. (1965). Optimum growth in an aggregative model of capital accumulation. *The Review of Economic Studies*, *32*(3), 233-240.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics*, *3*(1).

Castle, J. L., & Hendry, D. F. (2010). A low-dimension portmanteau test for nonlinearity. *Journal of Econometrics*, *158*(2), 231-245.

Castle, J. L., & Hendry, D. F. (2012). Automatic Selection for nonlinear Models. *System Identification, Environmetric Modelling and Control Systems Design*, 229-250.

Ciccone, A., & Jarocinski, M. (2010). Determinants of economic growth: Will data tell?. *American Economic Journal: Macroeconomics*, *2*(4), 222-246.

Chatterji, M. (1992). Convergence Clubs and Endogenous Growth. *Oxford Review of Economic Policy*, *8*(4), 57-69.

Clarke, D. (2013). GETS: Stata module to implement a General-to-Specific modelling algorithm. *Statistical Software Components*.

Cuaresma, J.C., & Doppelhofer, G. (2007). Nonlinearities in cross-country growth regressions: A Bayesian averaging of thresholds (BAT) approach. *Journal of Macroeconomics, 29*(3), 541-554.

De Luca, G., & Magnus, J. R. (2011). Bayesian model averaging and weighted average least squares: equivariance, stability, and numerical issues.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature*, 424-455.

Deijl, van der, W. (2013). *Data Mining: Vice or Virtue?*. Erasmus Institute for Philosphy and Economics Research Master Thesis.

Diamond, P. A. (1965). National debt in a neoclassical growth model. *The American Economic Review*, *55*(5), 1126-1150.

Doornik, J. A. (2009). Autometrics. In *in Honour of David F. Hendry*.

Doppelhofer, G., & Miller, R. I. (2004). Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach. *The American Economic Review*, *94*(4), 813-835.

Durlauf, S. N., & Johnson, P. A. (1995). Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics*, *10*(4), 365-384.

Durlauf, S. N., Johnson, P. A., & Temple, J. R. (2005). Growth econometrics. *Handbook of economic growth*, *1*, 555-677.

Durlauf, S. N., Kourtellos, A., & Minkin, A. (2001). The local Solow growth model. *European Economic Review*, *45*(4), 928-940.

Durlauf, S. N., & Quah, D. T. (1999). The new empirics of economic growth. *Handbook of macro economics*, *1*, 235-308.

Eberhardt, M., & Teal, F. (2011). Econometrics For Grumblers: A New Look At The Literature On Cross-Country Growth Empirics. *Journal of Economic Surveys*, *25*(1), 109-155.

Eicher, T., Papageorgiou, C., & Raftery, A. (2009). Determining growth determinants: default priors and predictive performance in Bayesian model averaging. *Journal of Applied Econometrics*.

Eicher, T. S., Papageorgiou, C., & Raftery, A. E. (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, *26*(1), 30-55.

Fernandez, C., Ley, E., & Steel, M. F. (2001a). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, *100*(2), 381-427.

Fernandez, C., Ley, E., & Steel, M. F. (2001b). Model uncertainty in cross-country growth regressions. *Journal of applied Econometrics*, *16*(5), 563-576.

Friedman, M. (1940). Review of Tinbergen (1939). *American Economic Review*, *30*(3), 657-60.

Geweke, J. (2010). *Complete and incomplete econometric models*. Princeton University Press.

Hendry, D. F. (2000). Econometrics: alchemy or science?: essays in econometric methodology.

Hendry, D. F., & Krolzig, H. M. (1999). Improving on 'Data mining reconsidered' by KD Hoover and SJ Perez. *The econometrics journal*, *2*(2), 202-219.

Hendry, D.F. and Krolzig, H.M. (2003).New Developments in Automatic General-to-specific Modelling. 379-419 in Stigum, B.P. (eds). *Econometrics and the Philosophy of Economics.* Princeton University Press.

Hendry, D. F., & Krolzig, H. M. (2004). We Ran One Regression*. *Oxford bulletin of Economics and Statistics*, *66*(5), 799-810.

Hendry, D. F., & Morgan, M. S. (1995) eds. *The foundations of econometric analysis.* Cambridge University Press.

Keuzenkamp, H. A. (1995). The econometrics of the Holy Grail—a review of econometrics: alchemy or science? Essays in econometric methodology. *Journal of Economic Surveys*, *9*(2), 233-248.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382-401.

Hollanders, D.A. (2011). Five methodological fallacies in applied econometrics. *Real-world economics review*, 57, pp. 115-126

Hoover, K. D., & Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *The Econometrics Journal*, *2*(2), 167-191.

Hoover, K. D., & Perez, S. J. (2000). Three attitudes towards data mining. *Journal of Economic Methodology*, *7*(2), 195-210.

Hoover, K. D., & Perez, S. J. (2004). Truth and Robustness in Cross-country Growth Regressions*. *Oxford bulletin of Economics and Statistics*, *66*(5), 765-798.

Kalaitzidakis, P., Mamuneas, T. P., Savvides, A., & Stengos, T. (2001). Measures of human capital and nonlinearities in economic growth. *Journal of Economic Growth*, *6*(3), 229-254.

Kennedy, P.E. (2002). Sinning in the Basesment: What are the Rules? The ten Commandments of Applied Econometrics. *Journal of Economic Surveys*, 16(4), pp. 569-589.

Keynes, J.M. (1939). Professor Tinbergen's Method. *The Economic Journal*, 49(195), pp. 558-577.

Keynes, J. M. (1940). On a method of statistical business-cycle research. A comment. *The Econoic Journal*, 154-156.

Keuzenkamp, H. A. (1995). The econometrics of the Holy Grail—a review of econometrics: alche my or science? Essays in econometric methodology. *Journal of Economic Surveys*, *9*(2), 233-248.

Koopmans, T. C. (1947). Measurement without theory. *The Review of Economics and Statistics*, *29*(3), 161-172.

Koopmans, T. C. (1965). On the Concept of Optimal Economic Growth," in The Econometric Approach to Development Planning. Amsterdam: North Holland.

Leamer, E. E. (1978). *Specification searches: ad hoc inference with nonexperimental data*. New York: Wiley.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31-43.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, *75*(3), 308-313.

Ley, E., & Steel, M. F. J. (1999). We have just averaged over two trillion cross-country growth regressions. *University of Edinburgh Discussion Paper*.

Ley, E., & Steel, M. F. (2007). Jointness in Bayesian variable selection with applications to growth regression. *Journal of Macroeconomics*, *29*(3), 476-493.

Ley, E., & Steel, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, *24*(4), 651-674.

Liu, Z., & Stengos, T. (1999). nonlinearities in cross-country growth regressions: a semiparametric approach. *Journal of Applied Econometrics*, *14*(5), 527-538.

Louçã, F. (2007). *The years of high econometrics: A short history of the generation that reinvented economics (ch.7)*. Psychology Press.

Lucas Jr, R. E. (1988). On the mechanics of economic development. *Journal of monetary economics*, *22*(1), 3-42.

Magnus, J. R., Powell, O., & Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics*, *154*(2), 139-153.

Masanjala, W. H., & Papageorgiou, C. (2004). The Solow model with CES technology: nonlinearities and parameter heterogeneity. *Journal of Applied Econometrics*, *19*(2), 171-201.

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. *Journal of Applied Econometrics*, *23*(5), 671-682.

Mayer, T. (2000). Data mining: a reconsideration. *Journal of Economic Methodology*, *7*(2), 183-194.

Moral-Benito, E. (2010). Model averaging in economics. *Documentos de Trabajo (CEMFI)*, (8), 1.

Norton, J. D. (2010). Challenges to Bayesian confirmation theory. *Philosophy of statistics*, *7*.

Quah, D. T. (1997). Empirics for growth and distribution: stratification, polarization, and convergence clubs. *Journal of economic growth*, *2*(1), 27-59.

Perez-Amaral, T., Gallo, G. M., & White, H. (2003). A Flexible Tool for Model Building: the

Relevant Transformation of the Inputs Network Approach (RETINA)*. *Oxford Bulletin of Economics and Statistics*, *65*(s1), 821-838.

Perez-Amaral, T., Gallo, G. M., & White, H. (2005). A comparison of complementary automatic modeling methods: RETINA and PcGets. *Econometric Theory*, *21*(1), 262-277.

Reichenbach, H. (1956). The Direction of Time. Berkeley, CA. University of California press.

Reiss, J. (2005). Causal instrumental variables and interventions. *Philosophy of science*, *72*(5), 964-976.

Reiss, J. (2008). *Error in economics: Towards a more Evidence-bas Methodology (ch.7)*. Abingdon: Routledge.

Romer, D. (2011). Advanced Macroeconomics (4th ed). McGraw-Hill.

Romer, P. M. (1986). Increasing returns and long-run growth. *The Journal of Political Economy*, 1002-1037.

Sachs, J., McArthur, J. W., Schmidt-Traub, G., Kruk, M., Bahadur, C., Faye, M., & McCord, G. (2004). Ending Africa's poverty trap. *Brookings papers on economic activity*, (1), 117-240.

Sala-i-Martin, X. X. (1996). The classical approach to convergence analysis. *The economic journal*, 1019-1036.

Sala-i-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, 178-183.

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

Salimans, T. (2012). Variable selection and functional form uncertainty in cross-country growth regressions. *Journal of Econometrics*.

Santos, C., Hendry, D. F., & Johansen, S. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, *23*(2), 317-335.

Serranito, F. (2004). Openness, growth and convergence clubs: a threshold regression approach. *CEPN, CNRS*.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics*, *70*(1), 65-94.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (Vol. 81). The MIT Press.

Tinbergen, J. (1939). Statistical Testing of Business Cycle Theories: Part II: Business Cycles in the United States of America, 1919-1932.

Tinbergen, J. (1940). On a method of statistical business-cycle research. A reply. *The Economic Journal*, 141-154.

Ulasan, B. (2011). Cross-country growth empirics and model uncertainty: An overview. *Economics Discussion Paper*, (2011-37).

Wooldridge, (2009). *Introductory Econometrics: a modern approach* (4th ed.). Canada: South-Western.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097-1126.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, *6*, 233-243.

Zeugner, S. (2012). Bayesian Model Averaging with BMS.